Design and application of structure-based pharmacophores for class A GPCRs

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann ,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op woensdag 20 Juni 2012
om 10.30 uur precies

door

Marinus Petrus Adrianus Sanders

geboren op 6 juli 1984
te Uden

Promotor

Prof. dr. J. de Vlieg

Co-promotor

Dr. S.B.Nabuurs

Manuscript-commissie

Prof. dr. S. Wijmenga

Prof. dr. L. Buydens

Prof. dr. P.A.J. Hilbers

This work was performed within the framework of Dutch Top Institute Pharma, project GPCR forum (D1-105).

Support by Merck Sharp & Dohme BV for this research and the publication of this thesis is gratefully acknowledged.

Title: Design and application of structure-based pharmacophores for class A GPCRs

ISBN/EAN: 978-90-804922-0-2

Waar een wil is,
is een weg

Table of contents

CHAPTER 1

General introduction

1.1 Drug discovery and development

The process of developing new medication is complex, expensive, time-consuming and full of risks. On average the development of a single new drug takes about 12 years and costs in the order of US $1 billion [1-3]. Despite the increasing investments in pharmaceutical research and development (R&D), there has been a steep rise in the attrition rate of drug candidates [4, 5]. This is currently one of the main challenges facing pharmaceutical industry as a whole. A global overview of the process from target discovery to market approval is shown in **Figure 1.1**.

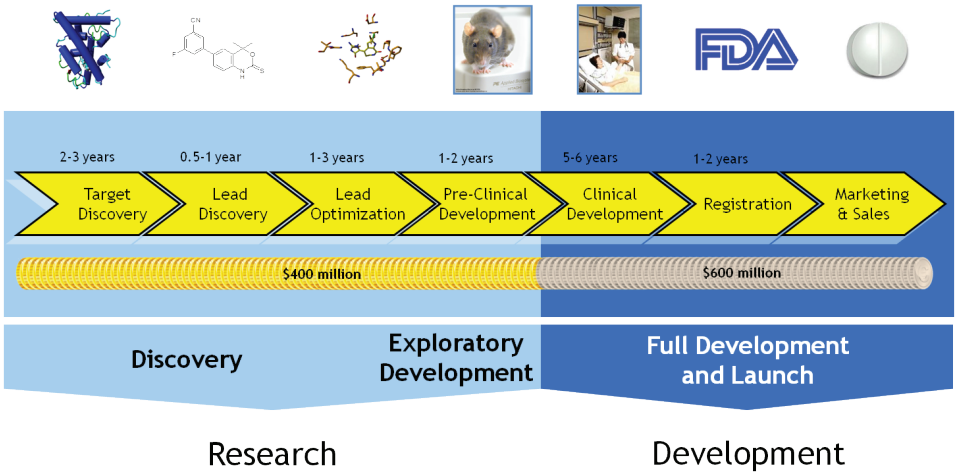


Figure 1.1: Drug discovery and development pipeline. The first essential step in the drug discovery pipeline is the target discovery, which involves the identification and early validation of disease-modifying targets. Subsequently a compound with the desired biological activity is identified in the lead discovery phase and its properties (potency, solubility, lipophilicity, metabolic stability, bioavailability, specific protein binding, and toxicity) are optimized in the lead optimization step. Explanatory development of leads is first pursued in animal models in the pre-clinical development phase, where the compounds safety profile is tested. Clinical development starts with phase I trials, in which the safety profile is checked in humans. These trials typically take 1-2 years and are performed in 20-80 healthy individuals. Subsequently, the efficacy and safety is tested in a small patient population of 100-300 patients which again takes 1-2 years. Next, a phase 3 trial is performed to test efficacy and safety in a large group of 1000-3000 patients. Finally, the drug is sent to regulatory authorities for review and approval and brought to the market upon successful completion.

In the target discovery phase, a disease modifying target has to be identified that can undergo a specific interaction with a drug to treat or diagnose the indication of interest. The vast majority of drug targets are proteins [6], which are either inhibited or activated via specific binding of a small molecule ligand. In the lead discovery phase compounds which modulate the biological activity of the selected target are identified by screening large compound collections *in vitro*. Subsequently in lead optimization, the identified bioactive compounds are chemically modified to improve potency and other desired drug-like properties, to for example allow safe oral administration. Next, compounds are

tested in preclinical development for the desired effect and to determine the toxicity profile in animal models. On average only 1 in 50 compounds passes this stage and is finally tested in human trials in the clinical development stage. Again only 1 out of 5 compounds in clinical development gets registered eventually and will be introduced to the market. As a result, the costs of drug development steeply rise after each step in the drug discovery process. It is therefore of crucial importance to keep the failure rate in the later stages of development as low as possible. Traditionally, this is done by assessing several protein targets and a large number of compounds to select the best target protein and lead compound. A relatively cheap solution is to use computational approaches which can be used to prioritize target proteins and drug candidate that have the desired properties to ultimately become a successful drug.

1.2 Pharmacophores

It is for example possible to characterize protein binding pockets by considering the relevant physicochemical properties and shape of the ligand binding pocket. Such descriptions are useful to assess the druggability of the target protein, but also to select compound sets with likely activity on the target and to chemically optimize a compound to improve activity or improve the properties of a compound while maintaining activity. An often used concept in this context is the pharmacophore. This is a three dimensional arrangement of chemical features necessary for biological activity [7-10] (**Figure 1.2**). These can be derived from either known active compounds or from the three dimensional target protein structure. Pharmacophores are especially useful for protein families for which only a limited number of high resolution protein structures are available.

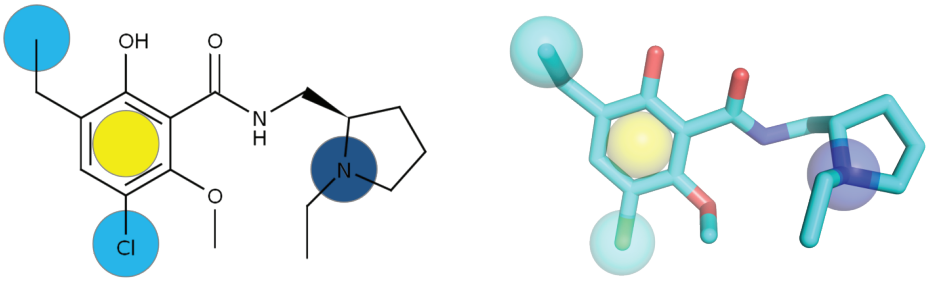


Figure 1.2: Schematic picture of a dopaminergic pharmacophore aligned to eticlopride [11] (left). Crystal structure pose of eticlopride as bound to the dopamine D3 receptor [12] with the corresponding 3D dopaminergic pharmacophore (right).

1.3 G protein-coupled receptors

G protein-coupled receptors (GPCRs), like other membrane proteins, are notoriously difficult to crystallize. Obtaining high-quantity and high-purity GPCR proteins is very challenging, because membrane proteins are typically produced in a heterogeneous environment by cells with varying glycosylation. Other challenges in the crystallization of membrane proteins relate to the flexible nature of constitutive active receptors and the

stabilization of the receptor structure as it is exposed to solvent.

In this thesis we describe the development and application of a pharmacophore modeling technique for the G protein-coupled receptor (GPCR) family. GPCRs comprise a large family [13, 14] of membrane proteins which are responsible for the signal transduction of endogenous signals into an intracellular response in many different physiological pathways [15]. As a result, GPCRs are effective drug targets for various diseases and of major interest to pharmaceutical companies. Of all drugs currently on the market approximately 25-50% interacts with a GPCR and new drugs targeting this protein family are continuously developed [15-20]. Most interesting from a drug development perspective is the class A family, also called rhodopsin-like, GPCRs. Characteristic for this family is that they contain relatively short N-termini. It is believed that the ligand binding pocket is located between the transmembrane region, which is shared between all GPCR's (**Figure 1.3**). This region resembles a barrel comprising seven structurally conserved α-helices that span the cell membrane in an anti-clockwise manner. Upon binding of an agonist, an activating molecule, to a GPCR, a conformational rearrangement of the intracellular segments of the transmembrane helices triggers a signaling cascade in the cytoplasm. In this process G-proteins and β-arrestins are considered as the two primary signal transducers.

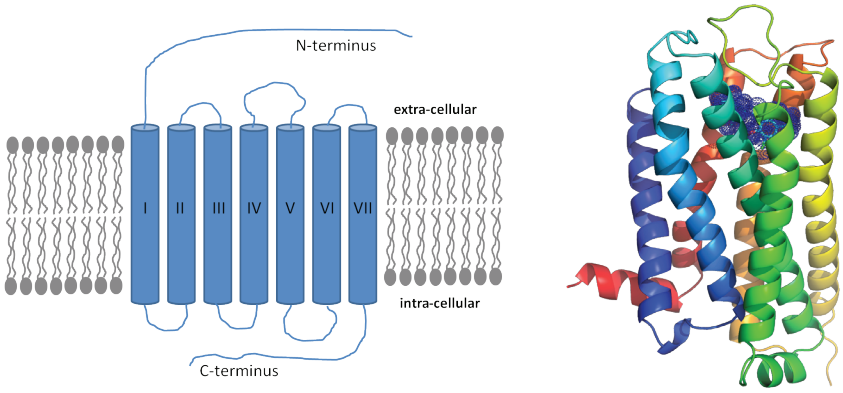


Figure 1.3: Schematic picture of a GPCR (left). Structure of the dopamine d3 receptor co-crystalized with eticlopride (right). The seven transmembrane helices are colored ranging from blue (TM I) to red (TM VII).

1.4 Outline of this thesis

The main focus of this thesis is the development and application of a software tool, named Snooker, which can be used to derive pharmacophores from protein structure models. A schematic outline of the work described in this thesis is depicted in **Figure 1.4**.

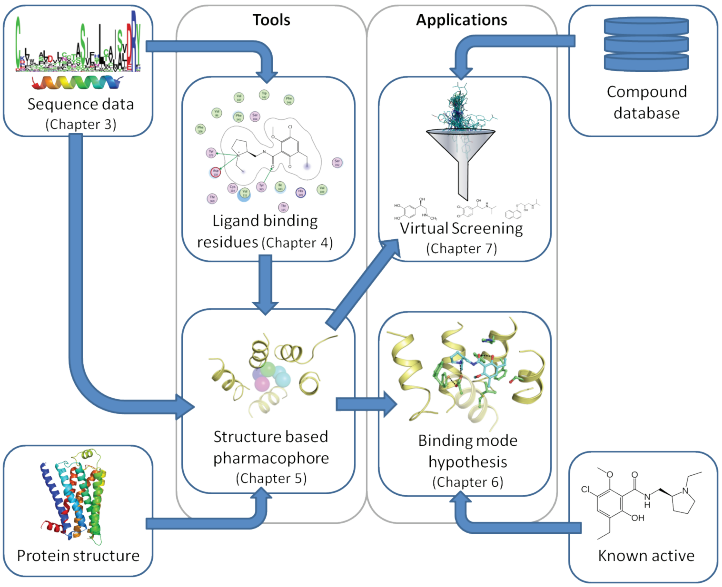


Figure 1.4: Outline of this thesis. Chapter two provides an overview of available algorithms to generate structure based pharmacophores. Chapter eight assesses the quality of pharmacophore search algorithms in terms of accuracy in virtual screening and binding mode hypothesis generation as applied to four different proteins.

Chapter 2 gives an overview of currently available methods to derive pharmacophores from protein structures and discusses possible applications of pharmacophores in drug design. **Chapter 3** describes the molecular class specific information system (MCSIS) GPCRdb. This database with annotated information about GPCR protein sequences of different species provides a good overview of the available data. In addition, it enables data mining to for example retrieve inter species differences, ligand binding effecting mutations or sequence patterns which can characterize certain subfamilies within the molecular class. In **chapter 4** we report ss-TEA, a method to predict ligand binding residue positions for GPCRs based on a multiple sequence alignment. **Chapter 5** describes how these ligand binding residue predictions are used to generate structure based pharmacophores with Snooker and how these can be applied to binding mode hypothesis generation and compound library design. **Chapter 6** reviews the successful ligand binding mode predictions produced by Snooker in combination with the flexible docking algorithm Fleksy, which were submitted to a global assessment on GPCR structure prediction. **Chapter 7** reports on a compound library design experiment in which Snooker was used in combination with a frequent substructure mining approach and also reports on the successful identification of new compounds for three different class A GPCR targets. Finally, in **chapter 8**, a benchmark study of different pharmacophore search algorithms is presented in which compound library enrichment and compound pose prediction is evaluated for four different protein targets.

References

1. Dickson, M. and J.P. Gagnon, *The cost of new drug discovery and development.* Discov Med, 2004. **4**(22): p. 172-9.
2. DiMasi, J.A., R.W. Hansen, and H.G. Grabowski, *The price of innovation: new estimates of drug development costs.* J Health Econ, 2003. **22**(2): p. 151-85.
3. Adams, C.P. and V.V. Brantner, *Spending on new drug development1.* Health Econ, 2010. **19**(2): p. 130-41.
4. Dimasi, J.A., *Risks in new drug development: approval success rates for investigational drugs.* Clin Pharmacol Ther, 2001. **69**(5): p. 297-307.
5. Mahajan, R. and K. Gupta, *Food and drug administration's critical path initiative and innovations in drug development paradigm: Challenges, progress, and controversies.* J Pharm Bioallied Sci, 2010. **2**(4): p. 307-13.
6. Smith, C., *Drug target validation: Hitting the target.* Nature, 2003. **422**(6929): p. 341-5
7. Ehrlich, P., Über den jetzigen stand der chemotherapie. Berichte der deutschen chemischen Gesellschaft, 1909. **42**(1): p. 17-47.
8. Kier, L.B., *Molecular orbital calculation of preferred conformations of acetylcholine, muscarine, and muscarone.* Mol Pharmacol, 1967. **3**(5): p. 487-94.
9. Kier, L.B., *The prediction of molecular conformation as a biologically significant property.* Pure Appl. Chem., 1973. **35**(4): p. 509-520.
10. Wermuth, C.G., et al., *Glossary of terms used in medicinal chemistry.* Pure Appl. Chem., 1998.
11. Klabunde, T. and A. Evers, *GPCR antitarget modeling: pharmacophore models for biogenic amine binding GPCRs to avoid GPCR-mediated side effects.* Chembiochem, 2005. **6**(5): p. 876-89.
12. Chien, E.Y., et al., *Structure of the human dopamine D3 receptor in complex with a D2/D3 selective antagonist.* Science, 2010. **330**(6007): p. 1091-5.
13. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
14. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
15. Filmore, D., *It's a GPCR world.* Modern Drug Discovery, 2004. **7**(11): p. 24-28.
16. Hopkins, A.L. and C.R. Groom, *The druggable genome.* Nat Rev Drug Discov, 2002. **1**(9): p. 727-30.
17. Klabunde, T. and G. Hessler, *Drug design strategies for targeting G-protein-coupled receptors.* Chembiochem, 2002. **3**(10): p. 928-44.
18. Crouch, M.F. and R.I. Osmond, *New strategies in drug discovery for GPCRs: high throughput detection of cellular ERK phosphorylation.* Comb Chem High Throughput Screen, 2008. **11**(5): p. 344-56.
19. Overington, J.P., B. Al-Lazikani, and A.L. Hopkins, *How many drug targets are there?* Nat Rev Drug Discov, 2006. **5**(12): p. 993-6.
20. Xiao, S.H., et al., *High throughput screening for orphan and liganded GPCRs.* Comb Chem High Throughput Screen, 2008. **11**(3): p. 195-215.

CHAPTER 2

From the protein's perspective: the benefits and challenges of protein structure-based pharmacophore modeling

*Marijn P.A. Sanders[1], Ross McGuire[2], Luc Roumen[3], Iwan J.P. de Esch[3], Jacob de Vlieg[1,4], Jan P.G. Klomp[5] and Chris de Graaf[*3]*

[1] Computational Drug Discovery Group, CMBI, Radboud University Nijmegen, Nijmegen, The Netherlands; [2] BioAxis Research BV, Berghem, The Netherlands; [3]Division of Medicinal Chemistry, LACDR, VU University Amsterdam, 1081 HV, Amsterdam, the Netherlands; [4]Netherlands eScience Center, Amsterdam, The Netherlands; [5]Lead Pharma Medicine, Nijmegen, The Netherlands

MedChemComm 2011, Accepted

Acknowledgments

The authors would like to thank Tina Ritschel and Sander Nabuurs for critical reading the manuscript. This work is supported by the Top Institute Pharma [project number D1.105: the GPCR Forum]. CdG is supported by the Netherlands Organization for Scientific Research (NWO) through VENI grant 700.59.408.

Abstract

A pharmacophore describes the arrangement of molecular features a ligand must contain to efficaciously bind a receptor. Pharmacophore models are developed to improve molecular understanding of ligand-protein interactions, and can be used as a tool to identify novel compounds that fulfil the pharmacophore requirements and have a high probability of being biologically active. Protein structure-based pharmacophores (SBPs) derive these molecular features by conversion of protein properties to reciprocal ligand space. Unlike ligand-based pharmacophore models, which require templates of ligands in their bioactive conformation, SBPs do not depend on ligand information. The current review describes the different steps in the construction of SBPs: i) protein structure preparation, ii) binding site detection, iii) pharmacophore feature definition, and iv) pharmacophore feature selection. We show that the SBP modeling workflow poses different challenges than ligand-based pharmacophore modeling, including the definition of protein pharmacophore features essential for ligand binding. A comprehensive overview of different SBP modeling and screening methods and applications is provided to illustrate that SBPs can be efficiently used for virtual screening, ligand binding mode prediction, and binding site similarity detection. Our review demonstrates that SBPs are valuable tools for hit and lead optimization, compound library design and target hopping, especially in cases where ligand information is scarce.

2.1. Introduction

The pharmacophore concept was first introduced in 1909 by Ehrlich who defined a pharmacophore as '*a molecular framework that carries (phoros) the essential features responsible for a drug's (pharmacon) biological activity*' [1]. Later, the pharmacophore concept was more precisely described by Kier [2] who stated that a drug must possess *'(a) those atomic features suitable for the requisite drug-receptor interaction phenomena and (b) the appropriate spatial disposition of these features necessary to bring about the required simultaneous or required sequential interaction events with the receptor*' [3]. Gund further updated Kier's definition and described a pharmacophore in 1977 as "*a set of structural features in a molecule that is recognized at a receptor site and is responsible for that molecule's biological activity*" [4]. Pharmacophores have proven to be extremely effective *in silico* filters in the search for bioactive molecules on several targets. Their use reduces the number of compounds and costs which have to be considered in large biophysical screenings [5-8]. In contrast to e.g. topological/2D ligand similarity searches which take a whole-structure, 'global,' view on the activity of molecules, pharmacophores focus on 'local' similarity and study the molecular determinants and their specific arrangement required for biological activity [9]. As a result they provide an explanation for the predicted activity of molecules. Ligand-based pharmacophores (LBPs), i.e., pharmacophore models derived from one or multiple active ligand(s), have been extensively used in the discovery and design of biologically active molecules [6]. Protein structure-based pharmacophores (SBP), derived from the three-dimensional (3D) structure(s) of one or more protein target(s), are receiving more and more attention in the past few years [6, 10]. One of the reasons for the rising interest in SBPs is the significant increase in high resolution protein structures. Currently more than 75.000 three-dimensional structures of biological macromolecules (mostly proteins) are deposited in the Protein Databank (9th August 2011)[11], leading to unprecedented understanding of these molecular drug targets. SBPs can be used as a tool to give insights into ligand-protein interactions and to enable large scale structural chemogenomics studies to identify new ligands for specific proteins (ligand profiling), or find new targets for specific ligands (target fishing)[12-16]. As such, there are three benefits of SBPs over LBPs: i) SBPs allow the identification of novel scaffolds, as they are less biased towards existing ligand chemotypes. ii) SBPs can be used to elucidate protein-ligand binding mode hypotheses within the protein structural framework [17-20], making SBPs suitable tools for structure-based ligand optimization. iii) SBPs lead to better understanding of ligand binding sites. These insights can for example be used to find ligands for orphan receptors [21] or to study ligand binding site similarities between different proteins to address cross-pharmacology [22] and suggest new targets for existing drugs [23].

Generally four different steps in the construction of SBPs can be distinguished: i) protein structure preparation, ii) binding site detection, iii) pharmacophore feature definition, and iv) pharmacophore feature selection (**Figure 2.1**). Protein structure preparation and

binding site detection are obviously specific aspects of SBP modeling. The protonation states of functional groups and conformation of the pharmacophore modeling template are clearly equally important determinants of LBPs and SBPs. For SBPs however, the possible variation in protonation states and conformations (at protein backbone, sidechain, or polar hydrogen atoms level) is in principle larger, simply because ligand binding sites (in most cases) contain more atoms than ligands do. Whereas the features included in SBPs are generally the same as in LBPs, the initial number of features in SBPs is generally higher. As a result, pharmacophore feature selection (an important step in both LBP and SBP modeling workflows) requires a different approach and poses different challenges in LBP and SBP modeling protocols. Structural alignment of different ligands can be used to identify essential features in LBPs, and experimental data such as structure-activity relationships (SAR) can be used to further emphasize specific features in LBPs. Selection of essential features in SBPs is however not straightforward, even when guided directly by experimental data such as site-directed mutagenesis studies or indirectly by including amino acid sequence based knowledge or ligand information. The larger number of conformational and spatial possibilities in SBPs make the selection of pharmacophore features more complex than in LBPs.

The current review gives a comprehensive overview of different methodologies to construct SBPs and describes the specific benefits and challenges (**paragraph 2.2.2**). Furthermore, representative applications of SBPs are provided to illustrate that SBPs can be efficiently used for virtual screening, ligand binding mode prediction, and binding site similarity detection (**paragraph 2.2.3**), Our review demonstrates that SBPs are valuable tools for hit and lead optimization, compound library design and target hopping, especially in cases where ligand information is scarce.

2.2. Protein Structure Based Pharmacophore Modeling Methodology

In principle, the availability of a three dimensional protein structure or model is the only absolute pre-requisite for SBP methods. If the target protein structure is not known a protein structure model may be generated by homology modeling with the use of the sequence and the three dimensional structure of a close homolog [24]. The subsequent steps in the SBP modeling workflow, protein preparation (**paragraph 2.2.1**), binding site detection (**paragraph 2.2.2**), pharmacophore feature definition (**paragraph 2.2.3**), and pharmacophore feature selection (**paragraph 2.2.4**) are outlined in **Figure 2.1** and discussed in this section. An overview and classification of the different SBP methods for binding site detection, feature definition and selection is presented in **Table 2.1**.

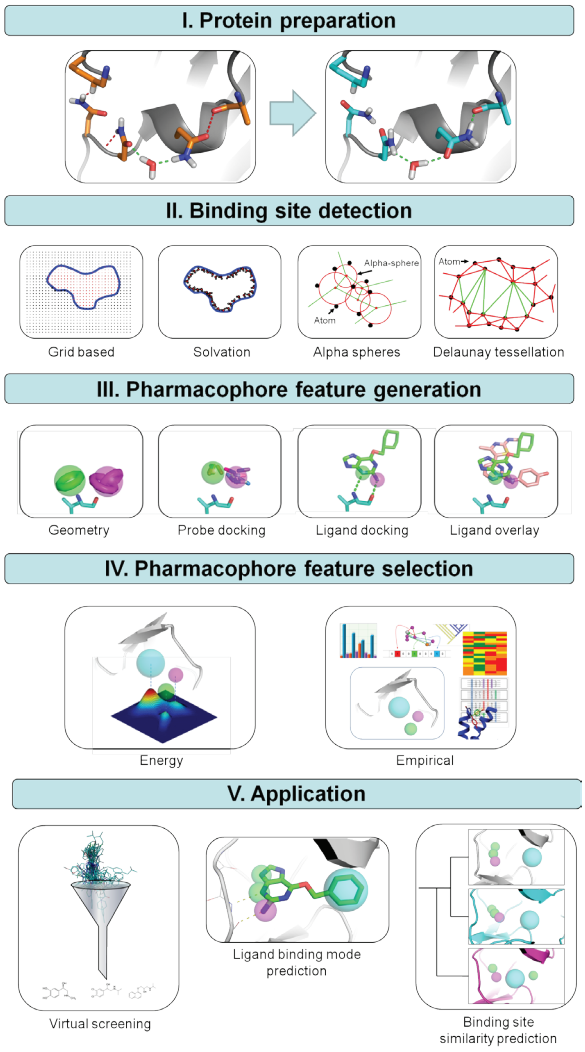


Figure 2.1: Methods for deriving structure based pharmacophores (SBPs). **I.)** The first step in deriving a SBP is the preparation of the protein structure. An example (PDB:2BNU) is shown with in green favorable interactions and unfavorable interactions in red. Three unfavorable interactions are observed in the original input file; an asparagine bumps into a lysine, a hydrogen of a neighboring asparagine clashes with the backbone and two partial negatively charged oxygens of an asparagine and backbone alanine are in close proximity. Flipping these asparagines removes all unfavorable interactions and gain one additional favorable interaction [65]. **II.)** Next, a cavity is defined using a grid-based, solvation, alpha sphere or Delaunay tessellation based approach. **III.)** Pharmacophore features are subsequently derived by calculation of spatially favorable positions for functional groups by using known interaction geometries, probe docking, or by docking one or more ligands and extracting this information from ligand poses. **IV.)** To increase pharmacophore specificity an energy or statistics based method is finally applied to reduce the number of pharmacophore features in the final pharmacophore hypothesis. **V.)** Three possible applications for pharmacophores are virtual screening, ligand binding mode prediction and binding site similarity prediction.

Table 2.1: Different structure based pharmacophore methodologies and their specifications.

Method/Software	Protein preparation	Binding site detection	Feature definition	Feature selection	Resolution[a]	Automation[b]R	Reference
Catalyst/Discovery Studio	Yes	Grid	Geometry	Energy	Medium	Medium	[50]
Chemogenomics (Kklabunde)	No	Fixed[c]	Fixed[c]	-	Low	High	[28]
FLAP	Yes	Grid	Geometry	Energy	Medium	Medium	[66, 67]
GBPM	Yes	Grid	Geometry	Energy[d]	Medium	Medium	[68]
HS-Pharm	Yes	Grid	Geometry	Empirical	Medium	Medium	[69]
LigandScout	Yes	-	Ligand	Energy	High	High	[70]
MOE	Yes	Alpha spheres	Ligand	Energy/ Manual	High	High	[52]
MUSIC	Yes	Solvation	Probe	Energy	Low	Medium	[47]
Pocket v2	Yes	Grid	Geometry	Energy	Medium	High	[71]
Schrodinger	Yes	Grid	Probe,Ligand	Energy	High	Medium, High	[72]
Snooker	No	Delaunay tesselation	Geometry	Empirical	Low	High	[73]
Sybyl	Yes	Grid,Solvation	Grid, Solvation	Knowledge, Energy/ Manual	Medium	Low	[49]

[a]High resolution methods require one high resolution structure from which all information is deduced. Medium resolution methods are able to combine different high resolution structures. Low resolution models do not require a high resolution structure, instead typically using interaction geometries, rotamer ensembles or simulations to cover a range of probable conformations.
[b]Highly automated methods can produce pharmacophores by the push of a button and have a fully integrated workflow. Medium automation requires a call to separate procedures. Low automation methods require the user to specify at which location or residue a feature should be positioned without giving much guidance except possibly the coloring of surfaces according to calculated properties.
[c]Method uses fixed position for pharmacophore features and positions them solely based on sequence information.
[d]Method requires ligand

2.2.1 Protein structure preparation

2.2.1.1 Protonation states

The first step in the derivation of a SBP is the preparation of the protein structure (**Figure 2.1**), including the consideration of non-protein groups (e.g. water molecules or cofactors), the determination of protonation states and positions of hydrogen atoms of protein residues, and the consideration of (alternative) protein conformations. Structures obtained by traditional X-ray crystallography and deposited in the PDB usually lack the positions of the hydrogens. These atoms are added according to basic pH dependent chemistry rules. Protonation states of aspartate, glutamate and histidine are typically solved by pKa predictions of the local structural environment while flipping of asparagines, glutamine and histidine side chains as well as tautomer optimization for histidine are performed to optimize the hydrogen bonding network [25]. An assessment of the accuracy of various methods for predicting hydrogen positions in protein structures has been reported by Forrest and Honig [26]. A protein preparation step is recommended since many structures deposited in the PDB were shown to contain errors [27]. An energy

minimization is optional to obtain the protein structure (or protein-ligand complex in case of a holo-structure) at the closest local minimum. Not all SBP methods require a correctly prepared protein structure (**Table 2.1**). Both the chemogenomics approach as described by Klabunde et al. [28] and Snooker [29] do not require high resolution 'cleaned' protein structures. The chemogenomics method uses fixed pharmacophore positions and needs only a correlation matrix between available pharmacophore features plus the sequence to create a pharmacophore. Snooker uses a structural template together with the sequence to build an ensemble structure which contains all likely rotamers of all residues. Since Snooker uses fixed rules for residue interaction geometries which are tautomer and protonation state independent, a protein preparation step is not required.

2.2.1.2 Protein conformations

Although LBP methods can in theory consider protein flexibility by aligning ensembles of molecular conformations, they typically assign primacy to the 'biologically active' ligand conformation and use only this to align multiple active molecules and deduce a pharmacophore [30-31]. However, a recent systematic study suggests that ligands rarely bind in their lowest calculated energy conformation [32], obviously making the problem of selecting biologically active conformations for constructing LBPs more complicated. Alternatively, many applications allow an externally derived conformation of an active molecule to be used as a template in the alignment of active molecules from which the pharmacophore is derived. In SBP modelling, it has been widely accepted that the conformational flexibility of proteins has to be taken into account. Flexible docking algorithms typically consider a number of proteins conformations and keep a limited number of residues flexible during the docking run [33-35] Carlson and Meagher developed an approach to incorporate protein flexibility in SBPs by generating pharmacophores from molecular dynamics (MD) simulations of protein-ligand complexes [36] based on clustering of the aligned pharmacophores of different snapshots. It is similarly possible to align the results of dockings, rendering this method suitable for the modeling of flexible proteins. Loving et al. [37] and Salam et al. [38] for example used Phase [39] to flexibly dock ligands/fragments and generate pharmacophore sites based on the docking solutions. They subsequently decomposed the Glide XP docking scores to energy terms for all individual probe fragments (hydrogen bond acceptors/donors, positive and negative ionizable groups and hydrophobic and aromatic interactions) and mapped these onto different the pharmacophore sites to rank the features. Using this approach Loving et al. reproduced the binding modes of 12 targets and recovered known actives from a database screen, while Salam et al. generated 30 successful reduced pharmacophore definitions and used them in enrichment studies. Geometry based methods also allow the generation of 'consensus' pharmacophores by overlaying protein structures and thus interaction maps. Molecular interaction fields (MIFs)[40], LUDI [41,42] and Snooker interaction maps generated from different models of the same

protein can be superposed to extract the most robust regions of interactions. Snooker by default uses the rotamer statistics as obtained from a rotamer library [43] to build an ensemble model for each protein.

2.2.2 Binding site detection

The second step in the SBP modeling workflow (**Figure 2.1**) is to define the location of the ligand binding site by the application of binding site detection algorithms. Retrospective studies show that these algorithms perform very well in the case of ligand bound crystal structures, correctly detecting up to 95% of the ligand binding pockets, but are less accurate for apo-structures [44]. Structure based binding site detection can be divided into energy based methods (incl. solvation methods) and geometry based (incl. grid based, alpha sphere, Delaunay tesselation) as summarized in **Figure 2.1** and **Table 2.1**. Energy-based algorithms, like PocketFinder [45], SuperStar [46], MUSIC [47], the solvation method of SiteID (Tripos) and SiteMap (Schrödinger) try to describe the local surface properties of a cavity by the simulation of solvent molecule interactions with the protein surface. SiteMap [48] uses a grid to sample binding site properties but is nevertheless fully based on interaction energy calculation. MUSIC [47] floods the pocket with probe fragments instead of solvent molecules and simulates the interactions of those fragments in short MD-simulations. Geometry based methods can be divided into two categories: discrete grid-based sampling and analytic methods like Delaunay tessellation and alpha sphere based methods. Grid-based methods available in SiteID (Tripos [49]), GRID [40] and in the binding site analysis module of InsightII (Accelrys [50]) typically select grid points close to but not overlapping with protein atoms, and define the pocket after a flood filling algorithm of grid points located in a cavity, where cavity grid points are defined as grid points with a substantial number of contacts with the protein. Snooker identifies the protein pocket by a Delaunay tessellation [51] of Cα-atoms and a calculated mean side chain atom, followed by the indexing of tetrahedra with at least one 'long' edge (**Figure2.1**). Delaunay tessellation implies the generation of an aggregate of space-filling irregular tetrahedral. These tetrahedra are deduced from a set of coordinates such that for each tetrahedron, the four vertices are on the circumspheres while no other vertices are inside the circumsphere. It is used to identify all tetrahedra which span large distances and cover potential ligand binding pockets. Finally, all indexed tetrahedra having a triangle in common are merged and the largest merged volume is defined as the pocket, while all non-indexed tetrahedra form the protein volume. SiteFinder (MOE [52]), PASS [53], SURFNET [54], LIGSITE [55], APROPOS [56], and CAST [57, 58] use alpha complexes to detect pockets. Alpha shapes are an extension of the convex hulls proposed by Edelsbrunner and Mucke [59]. Alpha spheres associated with 4 atoms can be generated from the simplices of a Delaunay tessellation (**Figure 2.1**). Solvent exposed alpha spheres and those corresponding to inaccessible areas (small spheres) are removed and the pocket is detected by aggregation of remaining nearby alpha spheres.

Two reviews summarizing and explaining ligand-binding site detection have been written by Prymula et al. [60] and Henrich et al. [61].

2.2.3 Pharmacophore feature definition

In the third step in the SBP flow scheme (**Figure 2.1**), pharmacophore features are derived from the co-crystallized ligand or from the ligand binding site (determined in the previous step) itself (**Table 2.1**). Protein structure based pharmacophore methods typically use geometric entities, such as spheres, vectors and planes with given attributes to characterize favorable interactions. The commonly used interaction types include H-bond acceptors, H-bond donors, positive and negative ionizable groups, lipophilic regions and aromatic rings. Their positions are set according to either positions of co-crystalized ligands or basic interaction geometry rules. Derivation of pharmacophores from ligands is ostensibly straightforward with features positioned at functional groups of the ligand. Ligand based pharmacophore modeling tools are generally similar although they may produce slightly different pharmacophores due to differences in feature definitions and algorithmic search strategies [62, 63]. An excellent review providing a detailed explanation of the available molecular alignment techniques has been written by Lemmen and Lengauer [64]. As each software package has differences in feature definition criteria, it is appropriate to use the same algorithm for both pharmacophore elucidation and pharmacophore searching, in order to ensure compatibility. If the structure of a protein-ligand complex is available, a pharmacophore can be constructed by positioning features at the functional groups of ligands (**Table 2.1**). Structure based pharmacophore derivation from apo structures is in contrast more challenging. The number of potential interacting residues in a ligand binding site is typically larger than the number of observed interactions in protein ligand complexes. Furthermore is it not straightforward to determine the optimal interaction geometry and there is no guarantee that ligands will interact at the predicted favorable sites of interaction. Some targets even have diverse ligand-binding modes and require a set of different pharmacophores to cover the interaction observed for all ligands [74-80]. Pharmacophore feature placement is therefore less accurate in structure based pharmacophore methods and tolerances of pharmacophore features are typically less strict. FLAP [66, 67], GBPM [68], Pocket V2 [71], Discovery Studio (Accelrys [50]), Sybyl (Tripos [49]) and Snooker [29] can create SBPs without the use of any ligand information and apply knowledge from residue based interaction geometries to predict likely interactions and their locations. MUSIC [47] and Schrodinger [37, 38] also do not require information about known actives and use a multiple copy simultaneous search (MCSS) method [81] to identify energetically favorable positions and properties of probes to generate pharmacophore features. In the MCSS method a protein's active site is filled with thousands of copies of organic functional groups which are allowed to energy-minimize onto the protein surface. Groups that minimize at the same location and bind most tightly are subsequently converted into pharmacophore

features. MOE [52] and LigandScout [70] at least require one known active ligand and a protein structure. Both can be used to generate a binding mode hypothesis and extract pharmacophore features from the modelled interactions. An improved pharmacophore can potentially be defined if multiple ligands are docked into the receptor active site and if only the conserved interactions are translated into pharmacophore features. Two recent comparative reviews on pharmacophore elucidation methods have recently been published by Luu et al. and Wallach. [82, 83].

2.2.4 Selection of essential pharmacophore features

To obtain valid binding mode hypotheses and subsequent compound library enrichment it is important to select only those features that correlate to biological activity (step 4 in the SBP modeling work flow (**Figure 2.1**)). Three different approaches can be defined to select essential pharmacophore features (**Figure 2.1, Table 2.1**): i) using interaction energy calculations (**paragraph 2.4.1**), ii) using protein-ligand interaction information (**paragraph 2.4.2**), and iii) based on analysis of amino acid sequence variation (**paragraph 2.4.3**). Finally, SBPs can be refined by training pharmacophore models with known active compounds (**paragraph 2.4.4**) and by complementing SBPs with shape restraints (**paragraph 2.4.5**).

2.2.4.1 Energy-based selection

Several SBP methods select pharmacophore features based on their (potential) interaction energy with ligands (**Table 2.1**). In cases where (the binding mode of) only one ligand is known, protein-ligand interaction energies can be estimated to prioritize interactions and discard those with small contributions to the overall binding energy. Methods using probe docking or simulation typically select features at positions where probes have high interaction energies. Here the possibility also exists to examine multiple protein structures obtained via experimental methods or homology modeling. Such structures can be overlaid after which 'hot spots' can be identified which represent conserved or highly favourable interactions. Most challenging is the selection of features after a geometry-based feature definition. Often, these methods place many pharmacophore features in the binding pocket and provide little information on which features to select. Sybyl [49] allows the user to manually pick the residues which correlate to binding activity and provides guidance via the calculation of surface properties. FLAP [66, 67] uses GRID [40] to generate molecular interaction fields (MIFs) which are condensed into discrete points representing the most favorable interactions. Pocket v.2 [71] uses a similar approach and generates a scored grid with the Pocket program to rank the protein-ligand interactions observed in the structure and utilizes this to automatically reduce the multitude of features to a reasonable number. Pocket v.2 also has the ability to suggest new binding spots besides the pharmacophore features already represented by a protein-ligand complex. Accelrys provides Ludi interaction maps [41] for H-bond donor,

H-bond acceptor and hydrophobic interactions. These are extracted from distributions of non-bonded contacts generated by a search through the Cambridge Structural Database [42] (which contains statistics about small molecule crystal structures) or generated by the application of interaction geometry rules which typically describe desired distances and angles between interacting pairs. Features are most likely in denser areas of those distributions. An alternative approach to GRID, LUDI and scored grids of the Pocket program is the extended electron distribution (XED)[84]. In contrast to the three methods previously described, XED generates field points based on a quantum orbital model. This enables the generation of multipoles for electronegative and atoms with Π orbitals resulting in a potentially more precise description of the desired sites of interaction.

2.2.4.2 Selection based on protein-ligand interaction information

If multiple ligands are available and binding mode hypotheses are generated, features which correlate to conserved protein-ligand interactions may be identified (Table 1). McGregor for example generated a pharmacophore for small molecule protein kinase inhibitors after extraction of conserved interactions in 220 kinase crystal structures [85]. Fingerprint methods like SIFt [86], SQUID [87] and FLAP [66] encode protein-ligand interactions in binary bitstrings and have been shown to be very useful for such analysis. The fingerprints typically contain information on the residue numbers and interaction type (HBA, HBD, positive ionisable, negative ionisable, hydrophobic, aromatic) observed in a protein-ligand complex. HS-Pharm [69] prioritizes cavity atoms that should be targeted for ligand binding, by training machine learning algorithms with atom based fingerprints of known ligand-binding pockets and was shown to have better enrichment curves for 2 out of 3 targets compared to docking algorithms [69]. An overview of pharmacophore methods classified according to the pharmacophore comparison method (alignment/ fingerprint based) can be found in a recent review by Luu et al. [82].

2.2.4.3 Selection based on variation of protein amino acids in the ligand binding site

Amino acid sequence variations can be used to evaluate the role of individual amino acid residues in the binding site and select those that are important for ligand binding. These data can be derived from single nucleotide polymorphisms (SNPs), site-directed mutagenesis studies, or sequence alignments of protein families. Snooker prioritizes cavity residues by analysis of a multiple sequence alignment (MSA) of a large set of homologous protein sequences. Klabunde [28] uses a set of 10 homology models and 3 X-ray structures to generate 35 single-feature pharmacophore elements associated with a sequence motif, the assumption being that certain sequence motifs at fixed positions are by definition important for ligand binding. Methods that use knowledge based prioritization of interactions (**Table 2.1**) are usually less dependent on the accuracy of the structures and models and are therefore better suited for protein families for which little structural data is available, like the G-protein coupled receptor (GPCR) family.

2.2.4.4 Training of SBPs with known actives

Known active compounds can be used to select specific (combinations of) pharmacophore features and to include shape restraints in the SBP model. Purely structure based pharmacophore models can be tested and optimized by searching a set of known active compounds and selecting only combinations of pharmacophore features which correlate to actives. Accelrys offers for example the Ligand Profiler Protocol to create a heatmap of ligands vs. pharmacophores. In this way, retrospective virtual screening studies can be used to identify pharmacophore features which discriminate known ligands from inactive (or decoy) molecules [69]. Snooker has been used to select compound sets by using ligand-based shape constraints based on poses of active compound in different sub-pharmacophores. The scoring function used by FLAP can be trained by supplying a set of known actives and known inactives and simultaneously minimizing the fraction of false positives and false negatives.

2.2.4.5 Shape restraints

Shape and volume are valuable concepts in drug design as outlined in a review by Nicholls et al. [88] Both concepts provide guidance for scaffold 'decoration' with chemical groups for virtual screening and lead optimization and can also be used in library design, ligand fitting, pose prediction, or active site description. The selectivity of pharmacophores can also be increased by the addition of shape restraints. Greenidge et al. [89] showed that the number of false positives can be decreased by a factor of 2-5 by the application of excluded volumes while the number of true positives remains nearly unchanged. Klabunde et al. [28] indicate that the enrichment they obtain is largely due to the addition of shape restraints to their initial pharmacophores. The program FLAP has the ability to describe the shape of the binding site by using shape probes in the GRID force field on which the fingerprints are based [66, 90]. Restraints can be added at positions occupied by the receptor, by space not occupied by a set of known actives or by setting a minimum shape similarity (volume overlap) to a reference compound. Using a shape restraint often ensures that especially large molecules with many interacting groups can only match the pharmacophore features in a conformation which is complementary with the protein binding site and do not match purely by increased probability. Rella et al. [91] showed the contribution of different shape restraints for angiotensin converting enzyme 2 inhibitors. First they generated a pharmacophore comprising 5 features (two hydrogen bond acceptors (HBA), two hydrophobic groups and one zinc binding group) and screened a compound library of 3.8M compounds of which they retrieved 1M compounds. By the addition of a shape restraint set at 130% of the co-crystalized ligand volume, a further reduction to 91000 compounds was achieved. Filtering of this set with 25 excluded volume spheres placed at positions occupied by the receptor gave 56000 compounds while only 38000 compounds were selected after the reduction in HBA tolerances to 1.3Å. A final shape filter of 110% and 100%, of the reference structure volume reduced the selection to

35000 and 16665 compounds, respectively. Seventeen compounds were selected based on high fit values as well as diverse structures and subjected to experimental validation in a bioassay. All these compounds showed an inhibitory effect on ACE2 activity. Some virtual screening methodologies are even entirely based on shape, like the ROCS method from OpenEye and Shape4 developed by Ebalunode et al. [92]. Both have been reported to perform well in terms of virtual screening [93, 94]. However to derive specific shape restraints the conformation and binding mode of at least one known active molecule is required and these methods have therefore limited compatibility with structure based pharmacophores.

2.3. Applications of SBPs

SBPs are developed to improve molecular understanding of ligand-protein interactions (**paragraph 2.3.1**). As already mentioned in **paragraph 2.1**, SBPs can be used for virtual screening (**paragraph 2.3.2**), ligand binding mode prediction (**paragraph 2.3.3**), and binding site similarity detection (**paragraph 2.3.4**). The lower part of **Figure 2.1** gives a pictorial description of three different uses of SBPs and an overview of relevant articles concerning the application of structure-based pharmacophores is presented in **Table 2.2**. The application papers discussed in this review are diverse with respect to generation method, target family, available data and used software and will be discussed in more detail in this section. The SBP models described in the current paragraph furthermore exemplify: 1) some of the potential advantages of SBPs over LBPs; 2) the discovery of novel scaffolds different from known chemotypes; 3) the elucidation of protein-ligand binding modes, 4) better understanding of ligand binding sites.

2.3.1 SBPs versus LBPs

While *comparative* ligand- and structure-based pharmacophore modelling studies are relatively scarce [95], more and more protocols are reported in which LBP and SBPs are combined [96-99]. This is in line with recent comparative virtual screening studies which show that ligand- and protein structure-based methods are complementary approaches in identifying different ligand chemotypes [99-101].

Thangapandian et al. [95] report a comparative study of ligand- and structure-based pharmacophores for the design of novel histone deacetylase 8 inhibitors. The LBP comprised 4 features retrieved 117 compounds of which 87 were active (corresponding to a ~8-fold enrichment over random picking), while. the SBP contained 6 features ,and retrieved 74 compounds of which 63 were active (corresponding to a comparable enrichment value of ~10). Kumar et al. [96] describe a method which combines LBPs and SBPs in order to identify additional interaction sites with c-Jun N-terminal kinase-3 that cannot be derived by ligand-based approaches alone. Using a training set of 21 c-Jun N-terminal kinase-3 inhibitors a LBP of 4 features was derived and used to construct a quantitative pharmacophore model to predict the affinities of 85 inhibitors with a

correlation coefficient r^2 of 0.846. Conserved hydrogen bond interactions with the hinge region were subsequently identified by docking of the 85 inhibitors into the kinase binding site and a SBP was derived which contained three additional features (2 donors and 1 acceptor) compared to the previously constructed LBP. Griffith et al. [97] and Singh et al. [102] have described protocols that combine the speed of LBPs with the ability of SBPs to predict ligand binding modes and to discover novel molecules.

Comparative retrospective virtual screening studies report similar [99] or somewhat higher [100, 101] overall enrichment for ligand-based methods compared to structure-based virtual screening approaches [101]. The performance of different ligand-based [103, 104] and docking-based methods [105] as well as the relative performance of ligand- vs. structure-based methods [99-101], can however be highly target dependent, justifying the use of both ligand-based and structure-based drug design techniques as complementary ligand discovery tools. Evers et al. [100] for example performed a retrospective virtual screeninig study on four biogenic amine-binding G-protein couped receptors (GPCRs) in which three ligand-based methods (LBPs [106, 107], 2D [108], and 3D similarity searches [109]) were compared with docking [110-112] in homology models. They showed that ligand-based methods outperform structure based methods, although structure-based methods still have satisfying enrichment factors (up to 60% of actives in the top-ranked 1% of the screened database). Krüger et al. [99] compared docking [113-116], 3D similarity searches [117] and 2D similarity searches [118] and reported almost equal enrichments. Hit lists obtained from different algorithms were however complementary and the combination of different approaches is likely to result in more (and more diverse) actives.

Table 2.2: Examples of structure based pharmacophore studies demonstrating the potential and wide variety of applications of SBPs for different protein targets

Target	Family	Input	Method	Prediction[a]	Result	Reference
AlaR	Racemase	X-ray	Overlay	VS	19 compounds selected for testing	[121]
C3AR1	GPCR	-	Chemoprints[b]	VS	4 new ligands found	[28]
various	Me-Lys binders	X-ray	Protein[c]	BS similarity	Similar sites identified	[139]
CDK2	Transferase	X-ray	Overlay	Model validation	recognition known ligands	[140]
CHK1	Transferase	X-ray	Overlay	VS	Enrichment > 9 fold	[122]
DHFR-TS	Synthase	X-ray	Complex	BM	correct prediction 6 ligands	[141]
DNMT1	Transferase	Model	Probe	BM + VS	Explanatory pharmacophore model derived	[142]
HIV-1 RT	Transcriptase	X-ray, NMR	Geometry	BM	HIV-1RT superligand generated	[138]
HtrA	Protease	Model	Geometry	VS	6 new ligands found	[120]
Kinase	Transferase	X-ray	Geometry	BS similarity	Relevant clustering of kinase families	[90]
Kinase Thrombin	Transferase Protease	X-ray	Geometry	BS similarity	Bio-isosters proposed for multiple targets	[143]
Kv1.5	Potassium channel	Model	Geometry	VS	19 new ligands found	[124]

Target	Family	Input	Method	Prediction[a]	Result	Reference
NK1R	GPCR	Model	Overlay	VS	1 new ligands found	[123]
NS3	protease/helicase/ NTPase	X-ray	Overlay	BM + VS	15 compounds selected for testing, ligand interacting residues identified	[135]
Renin	Angiotensinogenase	X-ray	Interaction	BM + VS	2 new ligands found	[136]
RSK2	Transferase	Model	Overlay	BM + VS	2 new ligands found	[144]
Thrombin	Protease	X-ray	Geometry	VS + scaffold hopping	Enrichment > 15 fold, successful scaffold replacement	[137]
TrXr, HIV-1N	Reductase, Integrase	X-ray	Geometry	VS	1 new ligand found	[119]
various	-	X-ray	Complex	VS	Average 40.1 fold enrichment	[38]
various	-	X-ray	Geometry	VS	Average 17 fold enrichment	[67]

[a] VS: virtual screening, BM: binding mode hypothesis generation, BS similarity: binding site similarity prediction
[b] Chemogenomics approach has fixed positions for features and positions features dependent on the protein sequence.

[d] Features are positioned on the protein (do not describe desired ligand features)

2.3.2 Virtual Screening (VS) for new ligands

Structure based pharmacophores (SBPs) are very well suited to combine the efficient screening methodologies of pharmacophores with structure information obtained from X-ray crystallization and NMR spectroscopy efforts. In this section 7 studies are described [28, 38, 67, 119-123] which use different methods (geometry (**paragraph 3.2.1**), probe (**paragraph 3.2.2**), complex **(paragraph 3.2.3**), overlay (**paragraph 3.2.4**), and chemogenomics (**paragraph 3.2.5**) based) to derive and select pharmacophore features and applied these to virtual screening (**Figure 2.1**).

2.3.2.1 VS using geometry based SBP

Tintori et al. used GRID to generate SBPs and showed that they could identify active molecules for thioredoxin reductase enzyme in a virtual screen and discover novel classes of active compounds able to inhibit complex formation between HIV-1 1N and viral host [119]. Pirard et al. used PASS (Putative Active Site with Spheres)[53] to identify the binding site and GRID to generate molecular interaction fields [124]. Manual selection of minima from the GRID energy maps and virtual screening with Unity [49] resulted in 19 novel potassium channel blockers. Cross et al. used FLAP to generate SBPs for 13 different targets extracted from the DUD (Directory of Useful Decoys) and showed enrichment values of approximately 17 fold over random at a false positive rate of 1% [67]. Significantly, they retrieved a variety of chemotypes demonstrating that lead-hopping and scaffold hopping into different chemical classes with SBPs is feasible [67, 125]. Löwer et al. used PocketPicker [126, 127] to extract binding pockets and generated interaction points complementary to the protein residues using Ludi rules [42, 128]. A pharmacophore was identified in regions of high interaction point density by the LIQUID program [129]. With this SBP six compounds (from 22 selected in silico hits) were identified that were able to block E-cadherin cleavage by HtrA [120].

2.3.2.2 VS using probe based SBP

Mustata et al. used molecular dynamics (MD) simulations to take protein flexibility into account in the construction of SBPs [121]. Ligbuilder [130] was used to generate property maps of the individual frames of the simulation of alanine racemase. Subsequently a dynamic pharmacophore was extracted using a MCSS approach [81] and the resulting SBP was used to identify compounds from ACD for experimental validation.

2.3.2.3 VS using complex based SBP

Salam et al. derived SBPs from 30 different protein-ligand complexes and used the energetic terms computed by the Glide XP scoring function to rank the importance of pharmacophore features. In a subsequent virtual screen they obtained enrichment values ranging from 3 to 100 (average 40) at 1% of the decoy database screened [38].

2.3.2.4 VS using overlay based SBP

Chen et al. derived a common feature pharmacophore by clustering multiple structure based pharmacophore features from different Chk1-ligand complexes in comparable binding modes [122]. This pharmacophore was used in combination with excluded volumes and shape constraints and showed an enrichment factor of >9 in a virtual screen. Evers et al. used the MOBILE method [131] to generate a NK1 receptor model that was able to accommodate known NK1 antagonists from structurally diverse classes. Using mutational data from literature and the features common to all known NK1 antagonists considered, they deduced a pharmacophore model which was used to select 7 compounds for biochemical testing, 1 of which showed affinity in the submicromolar range [123].

2.3.2.5 VS using chemoprint based SBP

Klabunde et al. used homology models and corresponding binding mode hypotheses to derive pharmacophore features associated with specific amino acid sequence motifs [28]. These so-called *chemoprints* were used to generate SBPs for the Urotensin-II receptor and Complement component 3a receptor 1 (C3AR1). Additional shape restraints were extracted from binding mode hypotheses of known active compounds and a virtual screen resulted in the successful retrospective identification of 81 Urotensin-II receptor ligands and in a prospective identification of 4 C3AR1 ligands.

2.3.3 Ligand binding mode prediction

In contrast to ligand-based pharmacophores, structure-based pharmacophores add to the understanding of the interaction of a (set of) ligand(s) with the protein. This is beneficial for affinity and selectivity prediction, defining Structure Activity Relationships (SAR), and hit optimization. SBPs can also be used for a complete exploration of the binding pocket and enable the targeting of residues that have not previously been utilized in

interactions with ligands. A prerequisite for all these studies is the correct prediction of the ligand binding. So far only few studies have shown that SBPs are capable to reproduce ligand binding modes of experimentally determined protein-ligand complexes. Loving et al. [37] showed that they could reproduce the binding modes observed in 12 different protein ligand complexes, while Sanders et al. used SBPs to correctly predict the binding modes of a set of beta-2-adrenergic agonists and antagonists/inverse agonists. SBPs can also been used in pharmacophore constrained docking and several studies have shown that pharmacophore constraints can significantly improve binding mode predictions and virtual screening enrichment [39, 132-134]. In this section several different methods to derive and use SBPs in binding mode prediction will be discussed.

2.3.3.1 Ligand BM prediction using geometry based SBP

Kaczor et al. have derived a SBP from a LUDI interaction map of the NS3 binding site [135]. This SBP was successfully used to identify new ligand interacting residues for this protein. Thangapandian et al. used Discovery Studio to generate LUDI maps of the crystal structure of renin with co-crystalized aliskiren and used the cluster pharmacophore tool to generate pharmacophore features [136]. Superimposition and analysis of two other structures with co-crystalized ligands overlaid on the aliskiren structure resulted in the selection of representative pharmacophoric features of catalytic importance. Ahlström et al. constructed a thrombin SBP model based on GRID molecular interaction fields (MIFs) for ligand scaffold replacements of active molecules [137]. A similarity search of curated scaffolds resulted in thrombin-derived scaffolds among the top solutions and docking of the entire molecules with replaced scaffolds showed feasible binding patterns. Griffith et al. used the Unity-3D module of Sybyl to generate a 'superligand' from different crystal and NMR structures to facilitate the design of structurally diverse inhibitors that interact with residues of HIV1-RT (mutants) in a novel manner [138].

2.3.3.2 Binding mode prediction using complex based SBP

Schormann et al. used LigandScout to extract pharmacophores for 8 different co-crystalized DHFR inhibitors from the same chemical series [141]. All individual complex derived pharmacophores contain 5 hydrophobic features, 2 donors and 1 negative ionizable feature. This pharmacophore was used together with docking to generate a ligand alignment and allow quantitative structure activity relation (QSAR) modeling with HASL [145] to predict affinity values for different inhibitors. Yoo et al. used the energy-optimized pharmacophore (e-pharmacophore)[37, 38] approach that is based on Glide XP energy terms to extract the most favorable sites of interaction from 14 docked inhibitors in human DNA methyltransferase 1 (hDNMT1) and derived an explanatory pharmacophore for hDNMT1 inhibitors [142].

2.3.3.3 Binding mode prediction using overlay based SBP

Zou et al. generated a SBP for CDK2 inhibition and reduced the number of features by selection of the top ranked 7 features as found in the 124 protein–ligand complexes [140]. They show that this most-frequent-feature pharmacophore encompasses previously reported ligand-based pharmacophore models. This pharmacophore was successfully used to discriminate CDK2 inhibitors from inactives and predict activities in retrospective virtual screening studies. Nguyen et al. constructed a homology model of the RSK2 N-terminal kinase domain and optimized the models for different classes of active molecules, mimicking the ligand induced structural changes of the ATP-binding site of RSK2 [144]. A common pharmacophore was subsequently constructed from 5 consistently recurring protein-ligand interactions.

2.3.4 Ligand binding site comparison

Many different methods to compare ligand binding sites based on pharmacophore or pharmacophore related properties have been developed in the past decade (for extensive reviews see Henrich et al. [61] and Kellenberger et al. [146]). Most methods, like FuzCav [147], CavBase [148] and PocketMatch [149] compare binding sites by assigning pharmacophore features directly to the protein. Campagna-Slater identified ligand binding sites that are chemically similar to known methyl-lysine binding domains using such SBP models [139]. The KRIPO method, developed by Ritschel et al., uses protein ligand interaction features derived from the ligand binding site to create 3D-pharmacophore fingerprints. This method has been successfully used to identify similar binding pockets and suggest structural modifications to ligands based on presumed bioisosteres [143]. Sciabola et al. used FLAP to compare protein binding sites and were able to cluster kinase protein families in a relevant manner, predict ligand activity across related targets and perform protein-protein virtual screening [90]. These methods are very well suited to identify proteins which can be selectively targeted and compounds which might have activity on proteins with pharmacophorically similar binding sites.

2.4. Conclusion

The current review describes the different steps in the construction of SBPs: i) protein structure preparation, ii) binding site detection, iii) pharmacophore feature definition, and iv) pharmacophore feature selection. SBP generation typically starts with a protein preparation step to correct and optimize the starting structure. Subsequently the ligand binding pocket can be defined by a pocket detection algorithm and several strategies can be used to convert the protein properties to ligand. A choice of method can be made depending on the resolution of the available protein structure or model and the availability of known active ligands and corresponding binding mode hypotheses. Geometric methods to derive SBPs are typically least restrictive and are together with the probe based method the only approaches which can be applied in the absence of known ligands. A real challenge in SBP design is the reduction of the typically high number of features in a structure based pharmacophore to only those features which are related to biological activity. Energy based methods to construct SBPs in these cases typically depend on the accuracy of the input structure and the binding mode hypothesis generated, while statistics based measures, such as those relying on protein-ligand complex or protein variance information, are generally more robust with a greater capability to deal with low resolution or low quality structures. The relatively simple concept of a pharmacophore makes it an attractive tool for various research applications. Several studies have described the successful use of SBPs for binding mode hypotheses generation, virtual screening and binding site similarity calculations, demonstrating that SBPs are valuable tools for hit and lead optimization, compound library design and target hopping, especially in cases where ligand information is scarce.

References

1. Ehrlich, P., Über den jetzigen stand der chemotherapie. Berichte der deutschen chemischen Gesellschaft, 1909. **42**(1): p. 17-47.
2. Kier, L.B., *Molecular orbital calculation of preferred conformations of acetylcholine, muscarine, and muscarone.* Mol Pharmacol, 1967. **3**(5): p. 487-94.
3. Kier, L.B., *The prediction of molecular conformation as a biologically significant property.* Pure Appl. Chem., 1973. **35**(4): p. 509-520.
4. Gund, P., *Three-dimensional pharmacophoric pattern searching.* Prog Mol Subcell Biol, 1977. **11**: p. 117-143.
5. Langer, T., *Pharmacophores in Drug Research.* Mol Inf, 2010. **29**: p. 470-475.
6. Leach, A.R., et al., *Three-dimensional pharmacophore methods in drug discovery.* J Med Chem, 2010. **53**(2): p. 539-58.
7. Sun, H., *Pharmacophore-based virtual screening.* Curr Med Chem, 2008. **15**(10): p. 1018-24.
8. Kubinyi, H., ed. *Success stories of computer-aided design*. Computer Applications in Pharmaceutical Research and Development, ed. S. Ekins. 2006, Wiley-Interscience: New York.
9. Eckert, H. and J. Bajorath, *Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches.* Drug Discov Today, 2007. **12**(5-6): p. 225-33.
10. Langer, T., *Pharmacophores in Drug Research.* Mol. Inf., 2010. **29**: p. 470-475.
11. Bernstein, F.C., et al., *The Protein Data Bank: a computer-based archival file for macromolecular structures.* J Mol Biol, 1977. **112**(3): p. 535-42.
12. Rognan, D., *Structure-Based approaches to target fishing and ligand profiling.* Molecular Informatics, 2009. **29**(3): p. 176-187.
13. Rognan, D., *Chemogenomic approaches to rational drug design.* Br J Pharmacol, 2007. **152**(1): p. 38-52.
14. Ekins, S., J. Mestres, and B. Testa, *In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling.* Br J Pharmacol, 2007. **152**(1): p. 9-20.
15. Ekins, S., J. Mestres, and B. Testa, *In silico pharmacology for drug discovery: applications to targets and beyond.* Br J Pharmacol, 2007. **152**(1): p. 21-37.
16. Meslamani, J. and D. Rognan, *Enhancing the Accuracy of Chemogenomic Models with a Three-Dimensional Binding Site Kernel.* J Chem Inf Model, 2011.
17. Pike, A.C., et al., *Structural aspects of agonism and antagonism in the oestrogen receptor.* Biochem Soc Trans, 2000. **28**(4): p. 396-400.
18. Pike, A.C., A.M. Brzozowski, and R.E. Hubbard, *A structural biologist's view of the oestrogen receptor.* J Steroid Biochem Mol Biol, 2000. **74**(5): p. 261-8.
19. de Graaf, C. and D. Rognan, *Selective structure-based virtual screening for full and partial agonists of the beta2 adrenergic receptor.* J Med Chem, 2008. **51**(16): p. 4978-85.
20. Moras, D. and H. Gronemeyer, *The nuclear receptor ligand-binding domain: structure and function.* Curr Opin Cell Biol, 1998. **10**(3): p. 384-91.
21. Evers, A., et al., *Virtual screening of biogenic amine-binding G-protein coupled receptors: comparative evaluation of protein- and ligand-based virtual screening protocols.* J Med Chem, 2005. **48**(17): p. 5448-65.
22. Milletti, F. and A. Vulpetti, *Predicting polypharmacology by binding site similarity: from kinases to the protein universe.* J Chem Inf Model, 2010. **50**(8): p. 1418-31.
23. Haupt, V.J. and M. Schroeder, *Old friends in new guise: repositioning of known drugs with structural bioinformatics.* Brief Bioinform, 2011.
24. Cavasotto, C.N. and S.S. Phatak, *Homology modeling in drug discovery: current trends and applications.* Drug Discov Today, 2009. **14**(19422931): p. 676-683.
25. Hooft, R.W., C. Sander, and G. Vriend, *Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures.* Proteins, 1996. **26**(4): p. 363-76.
26. Forrest, L.R. and B. Honig, *An assessment of the accuracy of methods for predicting hydrogen positions in protein structures.* Proteins, 2005. **61**(2): p. 296-309.
27. Joosten, R.P., et al., *A series of PDB related databases for everyday needs.* Nucleic Acids Res, 2011. **39**(Database issue): p. D411-9.
28. Klabunde, T., C. Giegerich, and A. Evers, *Sequence-derived three-dimensional pharmacophore models for G-protein-coupled receptors and their application in virtual screening.* J Med Chem, 2009. **52**(9): p. 2923-32.

29. Sanders, M.P., et al., *Snooker: A structure based pharmacophore generation tool applied to class A GPCRs.* J Chem Inf Model, 2011.
30. Kier, L.B., *Fundamental Concepts in Drug-Receptor Interactions*, ed. J.F. Danielli, J.F. Moran, and D.J. Triggle. 1970, London, New York: Academic Press.
31. Kier, L.B. and L.H. Hall, *Molecular Orbital Theory in Drug Research*. 1971, New York: Academic Press. 5.
32. Perola, E. and P.S. Charifson, *Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding.* J Med Chem, 2004. **47**(10): p. 2499-510.
33. Alonso, H., A.A. Bliznyuk, and J.E. Gready, *Combining docking and molecular dynamic simulations in drug design.* Med Res Rev, 2006. **26**(5): p. 531-68.
34. Carlson, H.A., *Protein flexibility and drug design: how to hit a moving target.* Curr Opin Chem Biol, 2002. **6**(4): p. 447-52.
35. Teodoro, M.L. and L.E. Kavraki, *Conformational flexibility models for the receptor in structure based drug design.* Curr Pharm Des, 2003. **9**(20): p. 1635-48.
36. Meagher, K.L. and H.A. Carlson, *Incorporating protein flexibility in structure-based drug discovery: using HIV-1 protease as a test case.* J Am Chem Soc, 2004. **126**(41): p. 13276-81.
37. Loving, K., N.K. Salam, and W. Sherman, *Energetic analysis of fragment docking and application to structure-based pharmacophore hypothesis generation.* J Comput Aided Mol Des, 2009.
38. Salam, N.K., R. Nuti, and W. Sherman, *Novel method for generating structure-based pharmacophores sing energetic analysis.* J Chem Inf Model, 2009. **49**(10): p. 2356-68.
39. Dixon, S.L., et al., *PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results.* J Comput Aided Mol Des, 2006. **20**(10-11): p. 647-71.
40. Goodford, P.J., *A computational procedure for determining energetically favorable binding sites on biologically important macromolecules.* J Med Chem, 1985. **28**(7): p. 849-57.
41. Bohm, H.J., *On the use of LUDI to search the Fine Chemicals Directory for ligands of proteins of known three-dimensional structure.* J Comput Aided Mol Des, 1994. **8**(5): p. 623-32.
42. Bohm, H.J., *The computer program LUDI: a new method for the de novo design of enzyme inhibitors.* J Comput Aided Mol Des, 1992. **6**(1): p. 61-78.
43. Lovell, S.C., et al., *The penultimate rotamer library.* Proteins, 2000. **40**(3): p. 389-408.
44. Schmidtke, P., et al., *Large-scale comparison of four binding site detection algorithms.* J Chem Inf Model, 2010. **50**(12): p. 2191-200.
45. An, J., M. Totrov, and R. Abagyan, *Pocketome via comprehensive identification and classification of ligand binding envelopes.* Mol Cell Proteomics, 2005. **4**(6): p. 752-61.
46. Verdonk, M.L., J.C. Cole, and R. Taylor, *SuperStar: a knowledge-based approach for identifying interaction sites in proteins.* J Mol Biol, 1999. **289**(4): p. 1093-108.
47. Carlson, H.A., et al., *Developing a dynamic pharmacophore model for HIV-1 integrase.* J Med Chem, 2000. **43**(11): p. 2100-14.
48. Halgren, T., *New method for fast and accurate binding-site identification and analysis.* Chem Biol Drug Des, 2007. **69**(2): p. 146-8.
49. Tripos. [cited 2011; Available from: http://tripos.com/.
50. Accelrys. [cited 2011; Available from: http://accelrys.com/products/discovery-studio/.
51. Delaunay, B., *Sur la sphere vide.* Otdelenie Matematicheskikh i Estestvennykh Nauk, 1934. **7**: p. 793-800.
52. Group, C.C. [cited 2011; Available from: http://www.chemcomp.com/software.htm.
53. Brady, G.P., Jr. and P.F. Stouten, *Fast prediction and visualization of protein binding pockets with PASS.* J Comput Aided Mol Des, 2000. **14**(4): p. 383-401.
54. Laskowski, R.A., *SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions.* Journal of Molecular Graphics, 1995. **13**(5): p. 323-30, 307-8.
55. Hendlich, M., F. Rippmann, and G. Barnickel, *LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins.* Journal of Molecular Graphics and Modelling, 1997. **15**(6): p. 359-63, 389.
56. Peters, K.P., J. Fauck, and C. Frommel, *The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria.* Journal of Molecular Biology, 1996. **256**(1): p. 201-13.

57. Liang, J., H. Edelsbrunner, and C. Woodward, *Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design.* Protein Science, 1998. **7**(9): p. 1884-97.
58. Binkowski, T.A., S. Naghibzadeh, and J. Liang, *CASTp: Computed Atlas of Surface Topography of proteins.* Nucleic Acids Res, 2003. **31**(13): p. 3352-5.
59. Edelsbrunner, H. and L. Mucke, *Three-dimensional alpha shapes.* ACM Transaction on Graphics, 1994. **13**(1): p. 43-72.
60. Prymula, K., T. Jadczyk, and I. Roterman, *Catalytic residues in hydrolases: analysis of methods designed for ligand-binding site prediction.* J Comput Aided Mol Des, 2011. **25**(2): p. 117-33.
61. Henrich, S., et al., *Computational approaches to identifying and characterizing protein binding sites for ligand design.* J Mol Recognit, 2010. **23**(2): p. 209-19.
62. Wolber, G., et al., *Molecule-pharmacophore superpositioning and pattern matching in computational drug design.* Drug Discov Today, 2008. **13**(1-2): p. 23-9.
63. Spitzer, G.M., et al., *One concept, three implementations of 3D pharmacophore-based virtual screening: distinct coverage of chemical search space.* J Chem Inf Model, 2010. **50**(7): p. 1241-7.
64. Lemmen, C. and T. Lengauer, *Computational methods for the structural alignment of molecules.* J Comput Aided Mol Des, 2000. **14**(3): p. 215-32.
65. Word, J.M., et al., *Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation.* J Mol Biol, 1999. **285**(4): p. 1735-47.
66. Baroni, M., et al., *A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application.* J Chem Inf Model, 2007. **47**(2): p. 279-94.
67. Cross, S., et al., *FLAP: GRID molecular interaction fields in virtual screening. validation using the DUD data set.* J Chem Inf Model, 2010. **50**(8): p. 1442-50.
68. Ortuso, F., T. Langer, and S. Alcaro, *GBPM: GRID-based pharmacophore model: concept and application studies to protein-protein recognition.* Bioinformatics, 2006. **22**(12): p. 1449-55.
69. Barillari, C., G. Marcou, and D. Rognan, *Hot-spots-guided receptor-based pharmacophores (HS-Pharm): a knowledge-based approach to identify ligand-anchoring atoms in protein cavities and prioritize structure-based pharmacophores.* J Chem Inf Model, 2008. **48**(7): p. 1396-410.
70. inteligand. [cited 2011; Available from: http://www.inteligand.com/ligandscout/.
71. Chen, J. and L. Lai, *Pocket v.2: further developments on receptor-based pharmacophore modeling.* J Chem Inf Model, 2006. **46**(6): p. 2684-91.
72. Schrodinger. [cited 2011; Available from: http://www.schrodinger.com/.
73. Sanders, M.P.A., et al., *Snooker: A structure based pharmacophore generation tool applied to class A GPCRs.* J Chem Inf Model, 2011. **51**(9): p. 2277-92
74. Al-Nadaf, A., G. Abu Sheikha, and M.O. Taha, *Elaborate ligand-based pharmacophore exploration and QSAR analysis guide the synthesis of novel pyridinium-based potent beta-secretase inhibitory leads.* Bioorg Med Chem, 2010. **18**(9): p. 3088-115.
75. Taha, M.O., A.G. Al-Bakri, and W.A. Zalloum, *Discovery of potent inhibitors of pseudomonal quorum sensing via pharmacophore modeling and in silico screening.* Bioorg Med Chem Lett, 2006. **16**(22): p. 5902-6.
76. Taha, M.O., et al., *Discovery of new potent human protein tyrosine phosphatase inhibitors via pharmacophore and QSAR analysis followed by in silico screening.* J Mol Graph Model, 2007. **25**(6): p. 870-84.
77. Taha, M.O., et al., *Pharmacophore modeling, quantitative structure-activity relationship analysis, and in silico screening reveal potent glycogen synthase kinase-3beta inhibitory activities for cimetidine, hydroxychloroquine, and gemifloxacin.* J Med Chem, 2008. **51**(7): p. 2062-77.

78. Taha, M.O., et al., *Combining ligand-based pharmacophore modeling, quantitative structure-activity relationship analysis and in silico screening for the discovery of new potent hormone sensitive lipase inhibitors.* J Med Chem, 2008. **51**(20): p. 6478-94.
79. Taha, M.O., et al., *Pharmacophore and QSAR modeling of estrogen receptor beta ligands and subsequent validation and in silico search for new hits.* J Mol Graph Model, 2010. **28**(5): p. 383-400.
80. Wallach, I. and R. Lilien, *Predicting multiple ligand binding modes using self-consistent pharmacophore hypotheses.* J Chem Inf Model, 2009. **49**(9): p. 2116-28.
81. Miranker, A. and M. Karplus, *Functionality maps of binding sites: a multiple copy simultaneous search method.* Proteins, 1991. **11**(1): p. 29-34.
82. Luu, T.T., N.O. Malcolm, and K. Nadassy, *Pharmacophore Modeling Methods in Focused Library Selection - Applications in the Context of a New Classification Scheme.* Comb Chem High Throughput Screen, 2011.
83. Wallach, I., *Pharmacophore inference and its application to computational drug discovery.* Drug Development Research, 2011. **72**(1): p. 17-25.
84. Cheeseright, T., et al., *Molecular field technology applied to virtual screening and finding the bioactive conformation.* Expert Opin. Drug Discovery, 2007. **2**: p. 131-144.
85. McGregor, M.J., *A pharmacophore map of small molecule protein kinase inhibitors.* J Chem Inf Model, 2007. **47**(6): p. 2374-82.
86. Deng, Z., C. Chuaqui, and J. Singh, *Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions.* J Med Chem, 2004. **47**(2): p. 337-44.
87. Renner, S. and G. Schneider, *Fuzzy pharmacophore models from molecular alignments for correlation-vector-based virtual screening.* J Med Chem, 2004. **47**(19): p. 4653-64.
88. Nicholls, A., et al., *Molecular shape and medicinal chemistry: a perspective.* J Med Chem, 2010. **53**(10): p. 3862-86.
89. Greenidge, P.A., et al., *Pharmacophores incorporating numerous excluded volumes defined by X-ray crystallographic structure in three-dimensional database searching: application to the thyroid hormone receptor.* J Med Chem, 1998. **41**(14): p. 2503-12.
90. Sciabola, S., et al., *High-throughput virtual screening of proteins using GRID molecular interaction fields.* J Chem Inf Model, 2010. **50**(1): p. 155-69.
91. Rella, M., et al., *Structure-based pharmacophore design and virtual screening for novel angiotensin converting enzyme 2 inhibitors.* J Chem Inf Model, 2006. **46**(2): p. 708-16.
92. Ebalunode, J.O., et al., *Novel approach to structure-based pharmacophore search using computational geometry and shape matching techniques.* J Chem Inf Model, 2008. **48**(4): p. 889-901.
93. Hawkins, P.C., A.G. Skillman, and A. Nicholls, *Comparison of shape-matching and docking as virtual screening tools.* J Med Chem, 2007. **50**(1): p. 74-82.
94. Nicholls, A., et al., *Molecular shape and medicinal chemistry: a perspective.* J Med Chem, 2010. **53**(10): p. 3862-86.
95. Thangapandian, S., et al., *Ligand and structure based pharmacophore modeling to facilitate novel histone deacetylase 8 inhibitor design.* Eur J Med Chem, 2010. **45**(10): p. 4409-17.
96. Kumar, B.V., et al., *Ligand-based and structure-based approaches in identifying ideal pharmacophore against c-Jun N-terminal kinase-3.* J Mol Model, 2011. **17**(1): p. 151-63.
97. Griffith, R., et al., *Combining structure-based drug design and pharmacophores.* Journal of Molecular Graphics and Modelling, 2005. **23**(5): p. 439-46.
98. Tan, L., et al., *Integrating structure- and ligand-based virtual screening: comparison of individual, parallel, and fused molecular docking and similarity search calculations on multiple targets.* ChemMedChem, 2008. **3**(10): p. 1566-71.
99. Kruger, D.M. and A. Evers, *Comparison of structure- and ligand-based virtual screening protocols considering hit list complementarity and enrichment factors.* ChemMedChem, 2010. **5**(1): p. 148-58.
100. Evers, A., et al., *Virtual screening of biogenic amine-binding G-protein coupled receptors: comparative evaluation of protein- and ligand-based virtual screening protocols.* J Med Chem, 2005. **48**(17): p. 5448-65.
101. von Korff, M., J. Freyss, and T. Sander, *Comparison of ligand- and structure-based virtual screening on the DUD data set.* J Chem Inf Model, 2009. **49**(2): p. 209-31.

102. Singh, N., et al., *A combined ligand-based and target-based drug design approach for G-protein coupled receptors: application to salvinorin A, a selective kappa opioid receptor agonist.* J Comput Aided Mol Design, 2006. **20**(7-8): p. 471-93.
103. Kirchmair, J., et al., *How to optimize shape-based virtual screening: choosing the right query and including chemical information.* J Chem Inf Model, 2009. **49**(3): p. 678-92.
104. Kooistra, A.J., et al., *Electron density fingerprints (EDprints): virtual screening using assembled information of electron density.* J Chem Inf Model, 2010. **50**(10): p. 1772-80.
105. Moitessier, N., et al., *Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go.* British Journal of Pharmacology, 2008. **153 Suppl 1**: p. S7-26.
106. Guner, O., O. Clement, and Y. Kurogi, *Pharmacophore modeling and three dimensional database searching for drug design using catalyst: recent advances.* Current Medicinal Chemistry, 2004. **11**(22): p. 2991-3005.
107. Guner, O.F., *Pharmacophore modeling in drug design: recent advances.* Curr Comput Aided Drug Des, 2011. **7**(3): p. 158.
108. Roy, K., *Topological descriptors in drug design and modeling studies.* Molecular Diversity, 2004. **8**(4): p. 321-3.
109. Lemmen, C., T. Lengauer, and G. Klebe, *FLEXS: a method for fast flexible ligand superposition.* J Med Chem, 1998. **41**(23): p. 4502-20.
110. Halperin, I., et al., *Principles of docking: An overview of search algorithms and a guide to scoring functions.* Proteins, 2002. **47**(4): p. 409-43.
111. Kitchen, D.B., et al., *Docking and scoring in virtual screening for drug discovery: methods and applications.* Nat Rev Drug Discov, 2004. **3**(11): p. 935-49.
112. Sotriffer, C. and G. Klebe, *Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design.* Farmaco, 2002. **57**(3): p. 243-51.
113. Friesner, R.A., et al., *Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy.* Journal of Medicinal Chemistry, 2004. **47**(7): p. 1739-49.
114. Jones, G., et al., *Development and validation of a genetic algorithm for flexible docking.* J Mol Biol, 1997. **267**(3): p. 727-48.
115. Jain, A.N., *Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine.* J Med Chem, 2003. **46**(4): p. 499-511.
116. Kramer, B., M. Rarey, and T. Lengauer, *Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking.* Proteins, 1999. **37**(2): p. 228-41.
117. Rush, T.S., 3rd, et al., *A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction.* J Med Chem, 2005. **48**(5): p. 1489-95.
118. Hessler, G., et al., *Multiple-ligand-based virtual screening: methods and applications of the MTree approach.* J Med Chem, 2005. **48**(21): p. 6575-84.
119. Tintori, C., et al., *Targets looking for drugs: a multistep computational protocol for the development of structure-based pharmacophores and their applications for hit discovery.* J Chem Inf Model, 2008. **48**(11): p. 2166-79.
120. Lower, M., et al., *Inhibitors of Helicobacter pylori protease HtrA found by 'virtual ligand' screening combat bacterial invasion of epithelia.* PLoS One, 2011. **6**(3): p. e17986.
121. Mustata, G.I. and J.M. Briggs, *A structure-based design approach for the identification of novel inhibitors: application to an alanine racemase.* J Comput Aided Mol Des, 2002. **16**(12): p. 935-53.
122. Chen, X.M., et al., *Structure-based and shape-complemented pharmacophore modeling for the discovery of novel checkpoint kinase 1 inhibitors.* J Mol Model, 2010. **16**(7): p. 1195-204.
123. Evers, A. and G. Klebe, *Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model.* J Med Chem, 2004. **47**(22): p. 5381-92.
124. Pirard, B., J. Brendel, and S. Peukert, *The discovery of Kv1.5 blockers as a case study for the application of virtual screening approaches.* J Chem Inf Model, 2005. **45**(2): p. 477-85.
125. Cross, S. and G. Cruciani, *Grid-derived structure-based 3D pharmacophores and their performance compared to docking.* Drug Discov Today, 2010. **7**(4): p. 213-219.
126. Weisel, M., E. Proschak, and G. Schneider, *PocketPicker: analysis of ligand binding-sites with shape descriptors.* Chem Cent J, 2007. **1**: p. 7.
127. Weisel, M., et al., *Form follows function: shape analysis of protein cavities for receptor-based drug design.* Proteomics, 2009. **9**(2): p. 451-9.

128. Bissantz, C., B. Kuhn, and M. Stahl, *A medicinal chemist's guide to molecular interactions.* J Med Chem, 2010. **53**(14): p. 5061-84.
129. Tanrikulu, Y., et al., *Scaffold hopping by "fuzzy" pharmacophores and its application to RNA targets.* ChemBioChem, 2007. **8**(16): p. 1932-6.
130. Wang, R., Y. Gao, and L. Lai, *LigBuilder: A multi-purpose program for structure-based drug design.* J Mol Model, 2000. **6**: p. 498-516.
131. Evers, A., H. Gohlke, and G. Klebe, *Ligand-supported homology modelling of protein binding-sites using knowledge-based potentials.* J Mol Biol, 2003. **334**(2): p. 327-45.
132. Claussen, H., et al., *The FlexX database docking environment--rational extraction of receptor based pharmacophores.* Curr Drug Discov Technol, 2004. **1**(1): p. 49-60.
133. Hindle, S.A., et al., *Flexible docking under pharmacophore type constraints.* J Comput Aided Mol Des, 2002. **16**(2): p. 129-49.
134. Joseph-McCarthy, D., et al., *Pharmacophore-based molecular docking to account for ligand flexibility.* Proteins, 2003. **51**(2): p. 172-88.
135. Kaczor, A. and D. Matosiuk, *Structure-based virtual screening for novel inhibitors of Japanese encephalitis virus NS3 helicase/nucleoside triphosphatase.* FEMS Immunol Med Microbiol, 2010. **58**(1): p. 91-101.
136. Thangapandian, S., et al., *Potential virtual lead identification in the discovery of renin inhibitors: application of ligand and structure-based pharmacophore modeling approaches.* Eur J Med Chem, 2011. **46**(6): p. 2469-76.
137. Ahlstrom, M.M., et al., *Virtual screening and scaffold hopping based on GRID molecular interaction fields.* J Chem Inf Model, 2005. **45**(5): p. 1313-23.
138. Griffith, R., et al., *Combining structure-based drug design and pharmacophores.* J Mol Graph Model, 2005. **23**(5): p. 439-46.
139. Campagna-Slater, V., et al., *Pharmacophore screening of the protein data bank for specific binding site chemistry.* J Chem Inf Model, 2010. **50**(3): p. 358-67.
140. Zou, J., et al., *Towards more accurate pharmacophore modeling: Multicomplex-based comprehensive pharmacophore map and most-frequent-feature pharmacophore model of CDK2.* J Mol Graph Model, 2008. **27**(4): p. 430-8.
141. Schormann, N., et al., *Structure-based approach to pharmacophore identification, in silico screening, and three-dimensional quantitative structure-activity relationship studies for inhibitors of Trypanosoma cruzi dihydrofolate reductase function.* Proteins, 2008. **73**(4): p. 889-901.
142. Yoo, J. and J.L. Medina-Franco, *Homology modeling, docking and structure-based pharmacophore of inhibitors of DNA methyltransferase.* J Comput Aided Mol Des, 2011. **25**(6): p. 555-67.
143. Ritschel, T., et al., *Extraction of useful bioisostere replacements from the PDB.* Journal of Cheminformatics, 2011. **3**(Suppl 1): p. 37.
144. Nguyen, T.L., et al., *Homology model of RSK2 N-terminal kinase domain, structure-based identification of novel RSK2 inhibitors, and preliminary common pharmacophore.* Bioorg Med Chem, 2006. **14**(17): p. 6097-105.
145. Doweyko, A.M., *Three-dimensional pharmacophores from binding data.* J Med Chem, 1994. **37**(12): p. 1769-78.
146. Kellenberger, E., C. Schalon, and D. Rognan, *How to measure the similarity between protein-ligand binding sites.* Curr. Comput.-Aided, 2008. **4**: p. 209-220.
147. Weill, N. and D. Rognan, *Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites.* J Chem Inf Model, 2010. **50**(1): p. 123-35.
148. Schmitt, S., D. Kuhn, and G. Klebe, *A new method to detect related function among proteins independent of sequence and fold homology.* J Mol Biol, 2002. **323**(2): p. 387-406.
149. Yeturu, K. and N. Chandra, *PocketMatch: a new algorithm to compare binding sites in protein structures.* BMC Bioinformatics, 2008. **9**: p. 543.

CHAPTER 3

GPCRDB: information system for G protein-coupled receptors

Bas Vroling[1], Marijn Sanders[1,2], Coos Baakman[1], Annika Borrmann[1], Stefan Verhoeven[2], Jan Klomp[2], Laerte Oliveira[3], Jacob de Vlieg[1,2], Gert Vriend[1]

1 CMBI, NCMLS, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands; 2 Department of Molecular Design and Informatics, MSD, Oss, The Netherlands; 3 Department of Biophysics, Escola Paulista de Medicina, Federal University of Sao Paulo, Brazil

Nucleic Acid Research, 2011;39; D309-319

Acknowledgements

We thank Bob Bywater for stimulating discussions, and Maarten Hekkelman and Tim te Beek for their support with computer science issues. Steve Pettifer, Dave Thorne and Phil McDermott are thanked for providing us with the Utopia Documents PDF reader and their support with connecting it to the GPCRDB .

This work was supported by the BioRange program of these Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

This work was supported by the EMBRACE project that is funded by the European Commission within its FP6 Programme, under the thematic area “Life sciences, genomics and biotechnology for health”, contract number LHSG-CT-2004-512092.

We thank TIPharma for financial support.

Abstract

The GPCRDB is a Molecular Class-Specific Information System (MCSIS) that collects, combines, validates, and disseminates large amounts of heterogeneous data on G protein-coupled receptors (GPCRs). The GPCRDB contains experimental data on sequences, ligand binding constants, mutations, and oligomers, as well as many different types of computationally derived data such as multiple sequence alignments and homology models. The GPCRDB provides access to the data via a number of different access methods. It offers visualization and analysis tools, and a number of query systems. The data is updated automatically on a monthly basis. The GPCRDB can be found online at http://www.gpcr.org/7tm/.

3.1 Introduction

G protein-coupled receptors constitute a large family of cell surface receptors. They regulate a wide range of cellular processes, including the senses of taste, smell, and vision, and control a myriad of intracellular signalling systems in response to external stimuli. GPCRs are a major target for the pharmaceutical industry as is reflected by the fact that more than a quarter of all FDA approved drugs act on a GPCR [1]. GPCRs are arguably one of the most-researched classes of proteins, but despite intensive academic and industrial research efforts over the past three decades, little is known about the structural basis of GPCR function. From about 350 genes that code for the non-olfactorial receptors in the human species [2], only about 30 are truly validated therapeutic targets [3], indicating this family's immense potential for future drug development. The fact that GPCRs can form homo-oligomeric and hetero-oligomeric complexes [4] has created a lot of new challenges and opportunities in the rational drug design process. In addition, a number of high-resolution crystal structures recently became available, providing new insights in receptor structure and function and giving the GPCR field a big stimulus.

Researchers who focus on one particular protein or a class of proteins are confronted with the fact that both the number and the size of databases are expanding at an ever-increasing pace. Although many databases like PDB [5], UniProtKB [6], KEGG [7], EMBL [8], GenBank [9], etcetera are invaluable for their research, for the average wet-lab scientist these databases are less suitable for gathering, integrating, and updating different types of data in an easy and efficient manner. Studies that involve carrying over information from one protein to the other seem simple at a first glance, however, the amount of data that needs to be collected from heterogeneous sources, converted to syntactic and semantic homogeneity, validated, curated, stored, and indexed, is enormous.

The GPCRDB is a data source that holds a large amount of heterogeneous data in a well-organized and easily accessible form. This data is validated, internally consistent, and updated regularly. In addition to being a one-stop GPCR resource, the data in the GPCRDB facilitates inferring new information using a wide spectrum of bioinformatics techniques. The GPCRDB is a paradigm for MCSIS technology [10, 11].

3.2 New features

The previous release of the GPCRDB [12] was almost entirely a static website, neither offering much dynamic content, nor possibilities for complex interactions or the use of computational tools. We addressed this problem by rewriting the entire system. The use of new tools and modern-day e-Science technologies has resulted in improved flexibility and greater user-friendliness. We have updated the methods for harvesting GPCR sequences, expanded the number of data types available, and added new tools and services to the GPCRDB. Nearly all of the functionality that is offered through the web interface is also available in the form of web services. This allows for the easy integration of the GPCRDB in custom built tools and scripts or in workflow management tools such

as Taverna [13] and Pipeline Pilot (http://www.accelrys.com/products/pipeline-pilot/). All pages now offer extensive context-sensitive help functionality, explaining what kinds of data are displayed and how to use the available interactive functionalities such as searching and computational tools.

3.3 Data content

The contents of the GPCRDB can be categorized in three classes: primary, secondary, and tertiary data. Sequence data, ligand binding constants, mutant information, structural data, and oligomer interactions make up the experimentally determined primary data. Data types such as multiple sequence alignments, homology models, correlation patterns, and entropy-variability data are inferred from these primary data, and fall in the category of secondary, or computationally derived data. Curator provided interpretations and other user help facilities make up the tertiary data category. **Table 3.1** shows a few vital statistics about the volume of the data content of the GPCRDB.

Table 3.1: Statistics for the September 2010 release of the GPCRDB.

Sequences	27045
Families (and multiple sequence alignments)	1270
Mutations	7703
Ligand binding data	12086
Protein structures	195
Homology models	22616
Residues	11290993
Species	1521
Oligomers	115

3.4 Primary data

3.4.1 Sequences

GPCR sequences are extracted from NCBI's NR database, which is a non-redundant protein sequence database with entries from a set of sequence repositories that include GenBank CDS translations, UniProtKB, and PDB. GPCR sequences are selected by classifying them against a database of Hidden Markov Models (HMMs). For each of the protein families in the GPCRDB a HMM is available. These HMMs are created from multiple sequence alignments (MSA) of the previous GPCRDB release. HMM files are created with the HMMER software package (http://hmmer.wustl.edu/).

As a first step in harvesting GPCR sequences we perform a BLAST search with all the HMM consensus sequences against the NR database. By using a very relaxed cut-off value we collect a large number of resulting hits, including many false positives. This step is necessary to limit the amount of sequences that will be used for the actual classification while ensuring a minimal false-negative rate. This step reduced the search space (for the September 2010 GPCRDB release) from about 11 million sequences to about 80

thousand. These hits are scored against the collection of HMMs to place them in the correct family or discard them as not being a GPCR sequence. Additional filter steps are applied such as filtering out sequence fragments and sequences that contain ambiguous amino acid characters, resulting in a final set of about thirty thousand sequences. The corresponding database entries of the selected sequences are retrieved with MRS [14] and additional data such as gene names and species information is extracted and stored. The GPCRDB holds for each sequence one principal access page. **Figure 3.1** shows an example of such a page.

The protein detail page (**Figure 3.1**) contains a panel that visualizes sequence annotations such as helix boundaries, cysteine bridges and glycosylation sites. These annotations are loaded in real time using the DAS distributed annotation system [15, 16] and are visualized by Dasty2 [17], resulting in always up-to-date annotations. We use the UniProt DAS server to retrieve sequence annotations.

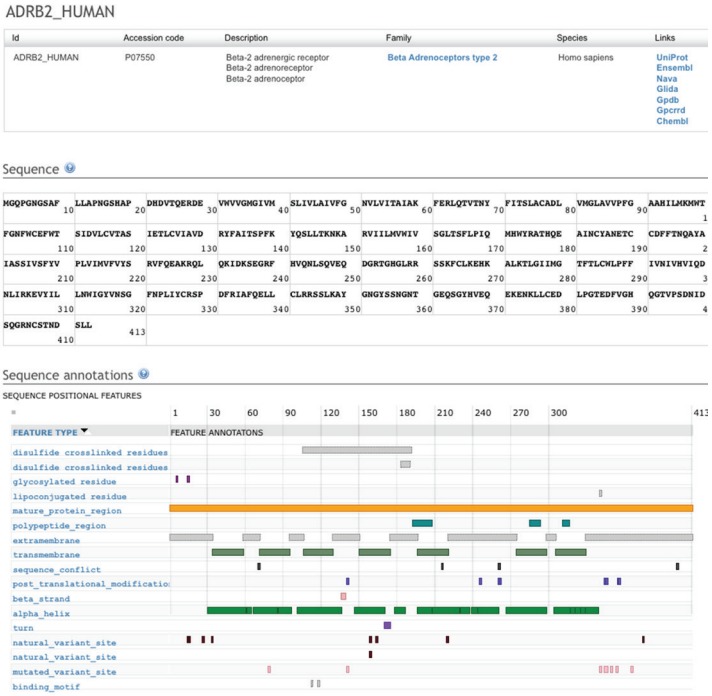


Figure 3.1: *Screenshot of the principal protein sequence page of the human beta-2 adrenoceptor. The top table contains details about the protein record and hyperlinks to the protein family browsing page and other data sources that contain information about this protein. The middle table holds the sequence in which each amino acid is linked to its own residue page. The bottom table holds annotations that are obtained in real time using the DAS (15) system.*

3.4.2 Ligand binding data

Ligand binding constants are available for a large number of GPCRs and are obtained from various sources. For each GPCR we provide links, if possible, to the ChEMBL [18] and GLIDA [19] databases. In addition, ligand-binding information that is obtained from collections from P. Seeman [20] and Organon N.V. [21] is available. Since ligand binding data is very hard to obtain from the literature we encourage academic and industrial researchers to submit their ligand binding data to the GPCRDB in order to make this data accessible to the scientific community.

3.4.3 Mutations

The GPCRDB contains a large number of well-annotated mutations obtained from different sources. We have two sets of mutations that were manually extracted from literature. Mutant data from the tinyGRAP database [22] contains references to scientific literature describing point mutations as well as insertions, deletions, and chimeric receptors. A collection of in-house manually extracted mutant data contains a few thousand point mutations and the effects of those mutations on the function of the receptor. We have extracted sentences from the papers that qualitatively describe the effects of these mutation and we have extracted quantitative data such as effects on ligand binding, expression, activation, or constitutive activity.

In addition to the two manually curated datasets we also have a large body of mutations that were extracted from the literature by the software package MuteXt [23]. A sentence describing the effects of the mutations is available for all mutations extracted by MuteXt.

3.4.5 Structures

Structures are obtained from the PDB. Links to structures that were re-refined in the PDB_REDO project [24] are included. We provide manually 'cleaned' monomers of the major GPCR PDB files that have been prepared for easy casual use by the life sciences community.

3.4.6 Oligomers

GPCR oligomerization has been an area of interest and controversy for many years. Recently there has been increasing evidence that both homo-dimers and hetero-dimers play a crucial role in GPCR signalling [25-27]. The GPCR-OKB [28] is a database that stores manually extracted computational and experimental information about GPCR oligomerization. Lists of protomers, experimental details and, where available, inferred oligomer interaction sites are available for all oligomers. This data has been fully integrated in the GPCRDB, making the information about both GPCR protomers and oligomers readily available.

3.5 Secondary Data

3.5.1 Multiple sequence alignments

Multiple sequence alignments (MSAs) are available for all families. MSAs are generated with WHAT IF [29] for all GPCR sub-families using hand-optimized sub-family specific profiles. Position-specific annotations such as secondary structure information and generalized residue numbers are stored in the profiles and are incorporated in the alignments. The general residue numbers are relative to the arbitrarily selected numbers for very conserved residues and motifs such as the well-known E/DRY and NPXXY motifs. Using a profile to align a GPCR sub-family allows for the mapping of the general residue numbers on the sequences that are being aligned. The result is that the residues in the TM domains, helix VIII and sections of the loops are labelled with a general residue number. For creating alignments of parent GPCR families we make use of these general residue numbers. For all the GPCRs that are being aligned we select all the general residue positions that the sub-families have in common and create the alignment by listing, for each sequence, the residues at the selected positions. GPCR parent family alignments are thus not built using standard alignment algorithms but are created by selecting residues that are likely to share the same position in the three-dimensional structure.

3.5.2 Homology models

Despite the recent publication of a number of GPCR structures the amount of structural information on GPCRs is still very limited. We have built structure models of all Class A receptors to at least partially fill this gap. Information extracted from the profile-based multiple sequence alignments is used to generate high-quality sequence-structure alignments between GPCR sequences and a number of experimentally determined structures. Based on these alignments homology models of all GPCRs have been automatically created and will be automatically updated as more sequences become available. Template structures are selected based on sequence identity and a number of structure quality criteria. The homology models are created with WHAT IF and YASARA [30]. Models will be automatically replaced when new structures become available that are better templates, with better being defined as either being solved with higher resolution data, or as having a higher percentage sequence identity.

3.5.3 Correlated mutation analyses

Correlated mutation analysis (CMA) is a technique that can find pairs of residues that remain conserved or mutate in tandem during evolution. Residues that show correlated behaviour in multiple sequence alignments are likely to be functionally related, and networks of those correlating residues indicate functional groups [31]. The rationale behind this analysis is that when a mutation occurs at a functionally important site, the protein either becomes functionally impaired or may acquire its original or a different function due to compensatory mutations at one or more other positions. Correlation scores are available in a number of different formats for all GPCR (sub)families.

3.5.4 Entropy and variability data

The amount of entropy and variability that is observed for a certain position in multiple sequence alignment tells something about the types of pressures exerted on that position during evolution [32-34]. Entropy-variability data is available for each multiple sequence alignment in the GPCRDB. We offer this data in tabular from, entropy-variability plots [35], and more advanced subfamily specific two-entropy plots [36] based on the original method described by Ye et al. [37] (**Figure 3.2**).

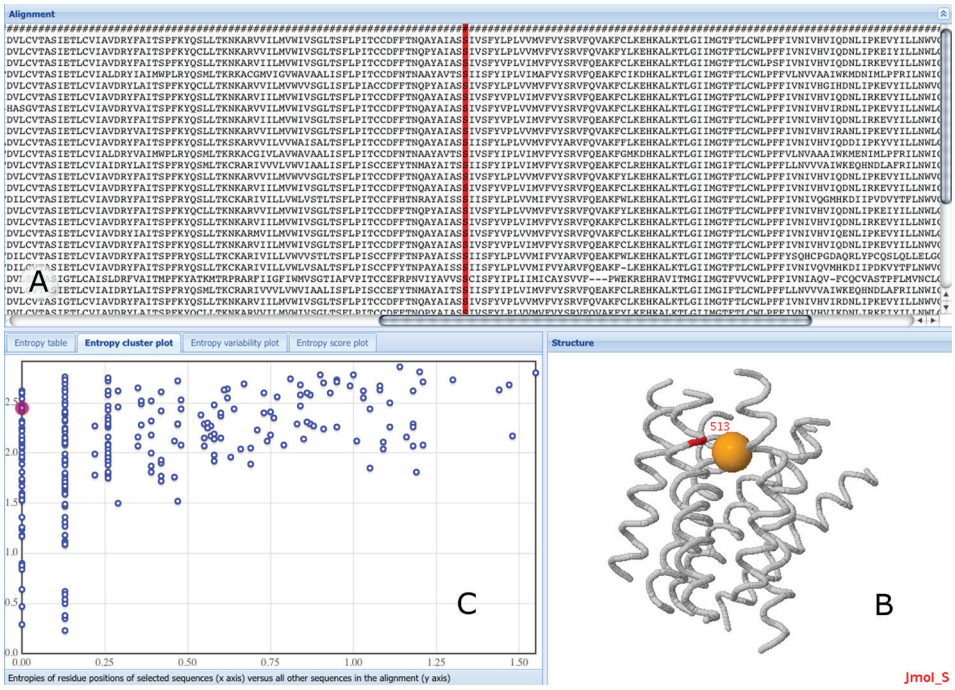


Figure 3.2: Screenshot of the interactive entropy and variability page. Residues are interlinked in all page elements, clicking results in highlighted selections. **A:** The multiple sequence alignment of the selected subfamily. **B:** the approximate location of this position in the 3D model of the transmembrane domain of class A GPCRs is shown in red and is annotated with general residue number information. The orange ball in the structure model indicates the approximate location of the assumed binding site for low molecular weight compounds of class A GPCRs. **C:** In this panel the user can choose among four display modes that describe the entropy and variability of all positions in the alignment; shown is the entropy cluster variant.

3.6 Tertiary data

3.6.1 Residue annotations

Residues in the GPCRDB are labelled with the original Oliveira et al. [38] numbering scheme as well as with residue numbers from the more recent Ballesteros-Weinstein scheme. Use of these general residue numbers allows for easy transfer of information between proteins. General residue numbers are available for all residues within

conserved structure elements. These include all transmembrane helices, helix VIII, and a few sections in the loops. For each residue with a general residue number a short description of its properties and interactions is available. These descriptions are based on a manual analysis of the currently available crystal structures.

3.6.2 Cytoscape networks

The GPCRDB provides cytoscape [39] network files for all GPCR families. These network files contain the proteins of a family with distances calculated from the family alignments. For all proteins the protein family information, species names and the amino acid types for all the residues annotated with a general residue number are available as attributes. This allows for complex analyses, such as colouring proteins by amino acids at a certain residue position to compare i.e. species or sub-type specific differences.

3.6.3 Mutation predictions

For all positions for which a general residue number is available we have investigated the most likely effects of mutations at these positions. Short fragments of text have been created that explain for each of these positions the likely effects of the mutation on structural and functional levels. References to key papers in which experimental evidence for these effects is available are included in the fragments. Examples of such effects are the loss of ligand binding affinity when mutating a residue in the ligand binding pocket, the loss of G protein binding when mutating residues at the G protein binding interface, and increased constitutive activity when residues are mutated a the interface between helix III and VI.

3.6.4 Workflows

We have created a number of Taverna workflows that use the web services of the GPCRDB as a starting point for users who want to programmatically access the GPCRDB. Workflows are available that use the GPCRDB BLAST service, create custom-built alignments and retrieve several different data types. The workflows and documentation are available via the myExperiment web portal [40] and are tagged with ‘GPCRDB’. We encourage researchers to share their own workflows via the myExperiment portal.

3.7 Data access

The GPCRDB provides fast and easy access to all its data and information. The GPCRDB does not merely lists available data, but all data types are fully integrated. For example, mutations are accessible via the protein detail pages but can also be found at the residue level. The same holds true for oligomer interaction interfaces, where details about these interactions are available via the oligomer pages, but also via the protein detail pages of interacting members, as well as via the pages of residues that are reported in the interaction. This tight data integration makes it a very efficient resource to use. The

GPCRDB's user interface allows the user to easily navigate from one data type to another and often suggests multiple routes to explore the data, thereby hopefully generating ideas and questions while the user navigates the system.

The four fundamental facilities to be provided by information systems are browsing, retrieval, query, and inferencing. These four types of access have been an integral part of the GPCRDB set-up from the beginning. The total redesign of the GPCRDB that has taken place the past few years has allowed us to add novel access facilities in all these four categories.

3.7.1. Browsing

The main way to access the data is via a hierarchical list of GPCR families, which is based on the pharmacological classification of GPCRs [41] (**Figure 3.3**). Users can traverse the GPCR family tree and view or download the data for a selected family. Available data types include multiple sequence alignments, entropy-variability analyses, and lists of family members. Alignments can be viewed in multiple ways. In addition to the classic HTML view, the GPCRDB offers an interactive multiple sequence alignment viewer (JalView [42], that can show additional information about the MSAs, supports a number of viewing and sorting options, and that can be used to generate phylogenetic trees. Residues for which mutation data is available are hyperlinked in the alignments to pages that contain more details about those mutations.

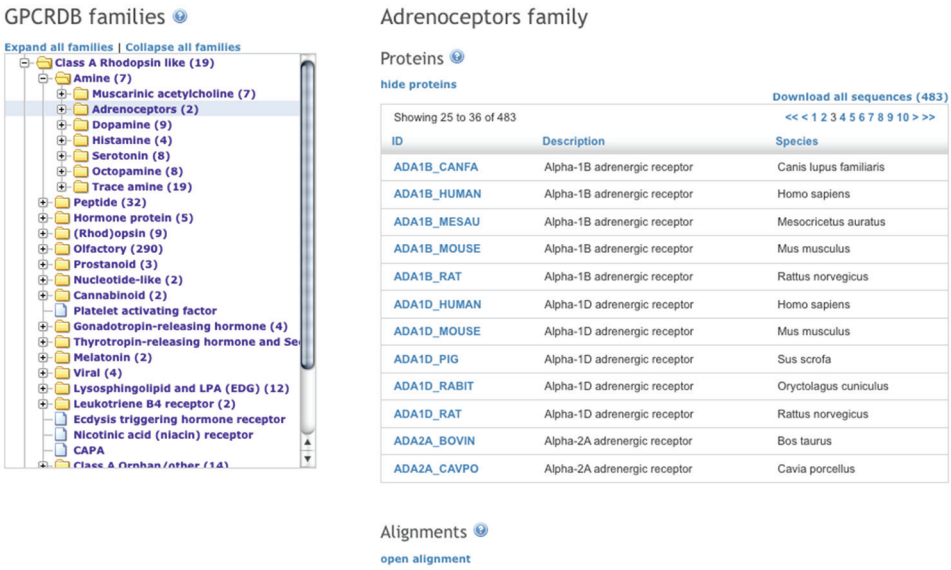


Figure 3.3: Screenshot of the GPCR family page. The GPCR family tree is shown on the left with the amine sub-family expanded. On the right-hand side the data for the selected family (adrenoceptors) is shown.

The pages that display data on individual proteins (**Figure 3.1**) contain a large amount of data and links to other data sources. **Table 3.2** lists the remote databases that have been indexed in the GPCRDB. Some of these remote data are actually most easily queried via the GPCRDB.

Table 3.2: *Non-GPCRDB data facilities that can be found through the GPCRDB.*

Database	Type of data	Address
GPCR-OKB (GPCR Oligomerization Knowledge Base)	Dimer information	http://data.gpcr-okb.org/gpcr-okb/
GPCRRD (GPCR Restraint Database)	Modelling restraints	http://zhanglab.ccmb.med.umich.edu/GPCRRD/
GLIDA (GPCR Ligand Database)	Ligand data	http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/
gpDB (G Protein Database)	G Protein data	http://biophysics.biol.uoa.gr/gpDB/
Uniprot	Protein information	http://www.uniprot.org/
Nava	Sequence variants	http://nava.liacs.nl/
ChEMBL	Ligand data	http://www.ebi.ac.uk/chembl/
Ensembl	Genomic information and annotations	http://www.ensembl.org

On the protein detail page the sequence is displayed and each residue is hyperlinked to its individual page where additional information is listed about that specific residue; residue numbers in multiple numbering schemes, available mutation data for that specific residue and mutations at equivalent positions, and reported oligomeric interactions.
Snake plots are available for all proteins in the GPCRDB (**Figure 3.4**). Residues for which mutations are available are hyperlinked from the snake plots to pages that contain details about those mutations.

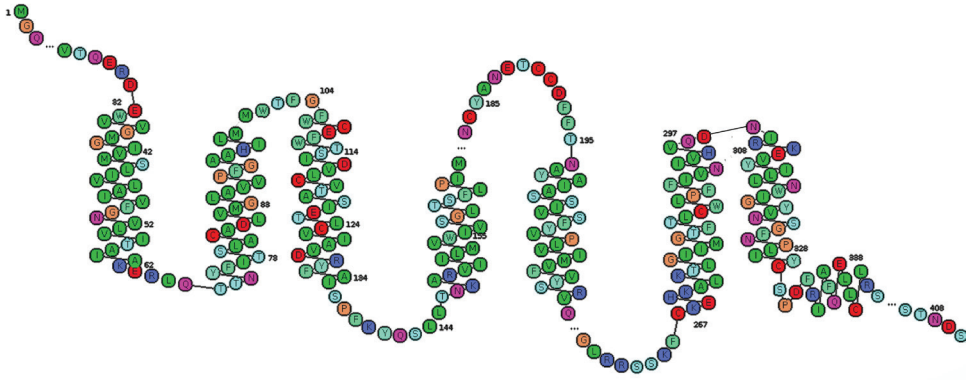


Figure 3.4: Snake plot of the human β2 adrenoceptor.

Other data types such as mutations, ligand binding constants and information about oligomerization states are displayed when available and links are provided to pages that contain more detailed information about those data (**Figure 3.5**).
The pages with mutation details contain links to the scientific literature and if available in that literature, the qualitative and quantitative data on the effects of the mutations is displayed. The oligomer detail page contains links to the GPCR-OKB and to the individual protomers in the GPCRDB. If certain residues are involved in the oligomer interaction, hyperlinks to individual residue pages are available.

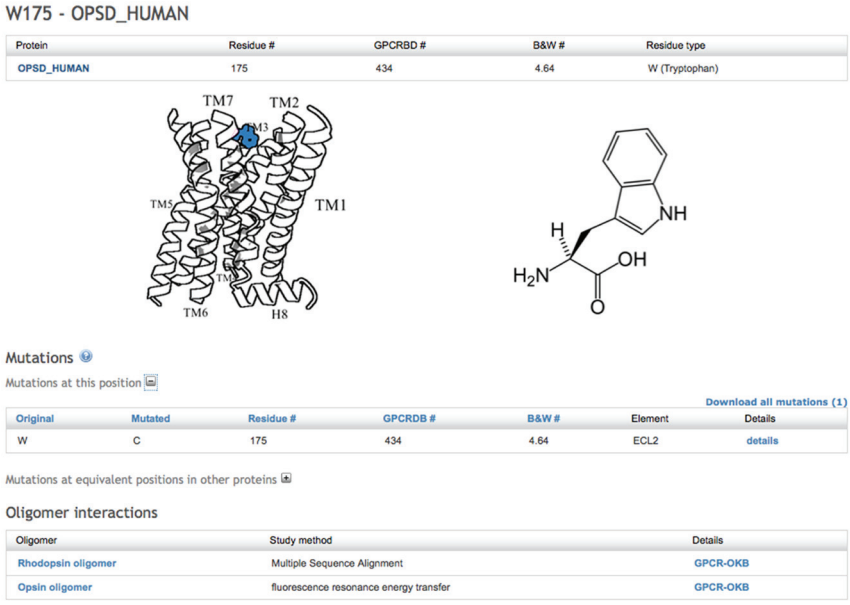


Figure 3.5: Screenshot of the detail page of residue W175 in OPSD_HUMAN. Residue numbers in different formats are shown, the approximate location of the residue is visually indicated and available mutations and oligomer data for this residue are listed.

3.7.2 Retrieval

Data can be retrieved via the web pages and via the web services. The web services offer very extensive retrieval possibilities, together allowing for the retrieval of all data types present in the GPCRDB. Subsets have been created for frequently requested data sets such as all human sequences and the human non-olfactory sequences. Protein family alignments can be downloaded in a number of different formats. Sequences, structures, ligand binding data and mutations can be downloaded from the protein detail pages. After querying the GPCRDB via the web pages, query result sets can be downloaded in FASTA format. A complete copy of the GPCRDB is freely available for in-house usage by academic and industrial researchers alike.

3.7.3 Query

Users can query the GPCRDB via a number of different search pages. Identifiers, genes, species, descriptions, and protein family names can be used to search for GPCRs (**Figure 3.6**). There are a number of filters available to limit the search results. Users can indicate whether only GPCRs should be shown for which mutations, structures, oligomers, or ligand binding data are available.

Mutations can be found via the mutant search page, where one can search by residue number (multiple numbering schemes are available) and/or residue types. The GPCRDB offers a BLAST service that allows users to BLAST their sequence against the sequences in the GPCRDB.

All search options and the BLAST services are available via the web interface and as web services. A full SQL search facility will be made available in the near future to allow for complex queries and analyses.

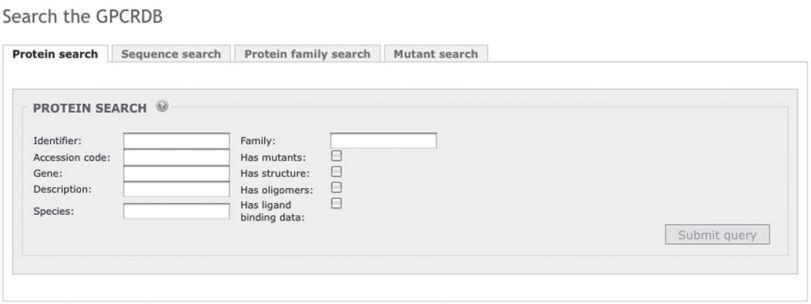


Figure 3.6: The protein search page.

3.7.4 Inferences

The amount of available GPCR related data is too large for a human to grasp and disseminate. The GPCRDB contains a series of inference engines that determine interesting correlations between the data, while a series of software tools help the user with data reduction and abstraction.

3.7.4.1 Building alignments

The GPCRDB offers the possibility to create custom-made alignments. The alignments are created by using the procedure that is used for the parent GPCR families as discussed earlier. Users can select the proteins and residue positions that should be aligned, allowing for the creation of e.g. an alignment of all binding pocket residues for a selection of proteins (**Figure 3.7**). The custom-built alignments are available for download and users can directly interact with the alignments using JalView.

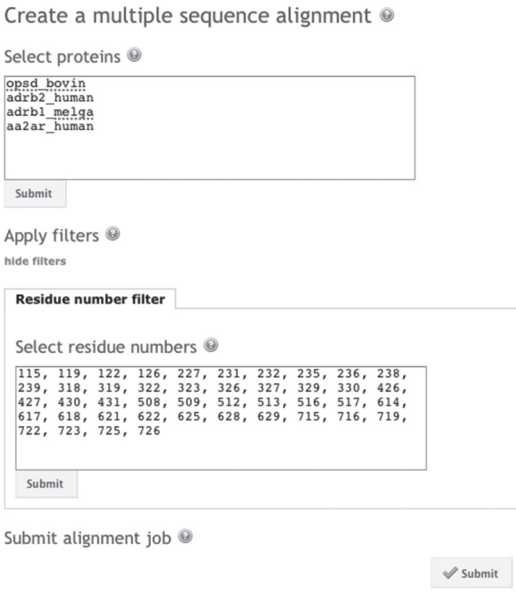


Figure 3.7: *A list of proteins and an optional list of residue positions can be used to generate custom alignments. In this figure we have selected a number of proteins for which crystal structures are available. The GPCR binding pocket residue positions as proposed by Gloriam et al. [43] are used. The result will be an alignment of all pocket residues of the selected proteins.*

3.7.4.2 Predicting the effects of mutations

We have started to create a service where users can predict the effects of a point mutation. As for now, predictions are mainly based on human knowledge that is stored in a computer readable format. This information is combined with a number of simple analyses on a homology model of the receptor being mutated, such as looking for steric clashes and helix disruptions. Results are presented as text-fragments that explain the effects of the mutation on the structure. Care has been taken to ensure that the results are presented in a life scientist friendly manner. The text contains references to literature and is enriched with figures and animations of the mutation and its surrounding environment. In the near future more intelligence will be added to the software, such as incorporating the quantitative data from the mutations extracted from literature and ligand binding information.

3.7.4.3 Analysis of entropy derived patterns

We offer a page where users can interactively analyze the protein family alignments (**Figure 3.2**). Plots displaying entropy and variability scores are displayed with a 3-dimensional model and a multiple sequence alignments. Residues are linked in the three page elements, so that clicking on a residue position in the multiple sequence alignment will highlight the residue position in the structure as well as in the entropy and variability plots. This offers researchers a very intuitive way of looking at conservation scores, even at the subfamily or receptor level, and relating those scores to the 3D structure. In combination with above mentioned accessible data site directed mutagenesis candidate selection, homology modeling and ligand binding hypotheses generation can be performed.

3.7.4.4 Annotating scientific literature

We have developed a new interface for the GPCR data in the form of a GPCR-specific PDF reader [44]. This reader can annotate scientific literature on GPCRs on the fly, providing users with context sensitive data from the GPCRDB (**Figure 3.8**). This software is available upon request and will be made freely available at the day of publication of this article.

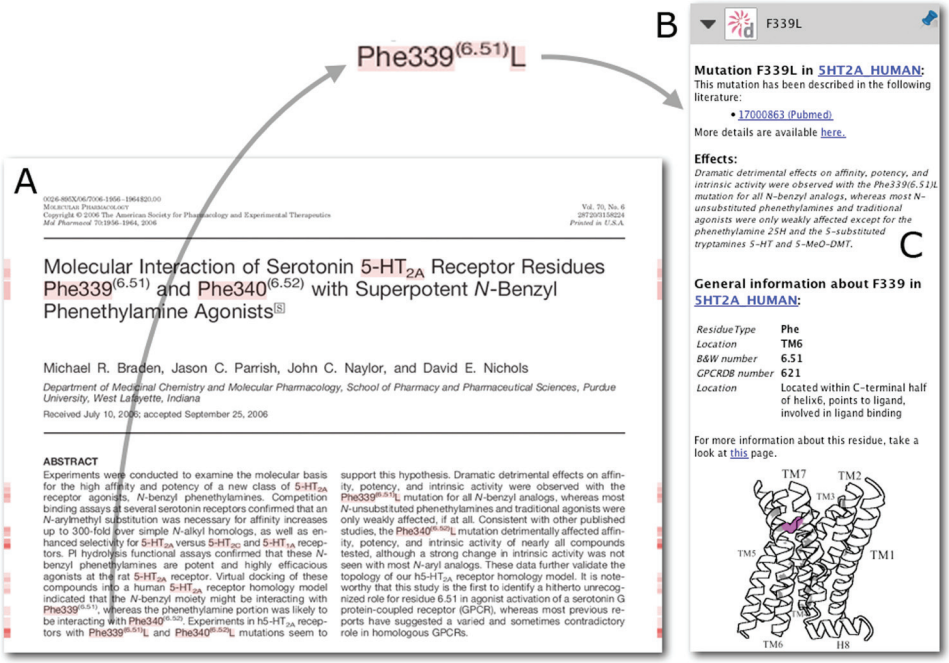


Figure 3.8: An impression of the PDF reader (Utopia Documents [45], Utopia Documents-GPCRDB (in preparation)) interface to the GPCRDB data. A: A scientific paper [46] is shown that is annotated by the GPCRDB. Annotations are available for all the highlighted words. B: An example of such an annotation (the mutation F339L) is displayed. C: A short, manually extracted description of the effects of this mutation is included.

3.8 Implementation

The data in the GPCRDB is stored in a PostgreSQL (http://www.postgresql.org/) relational database. The web service interface is developed with the Apache CXF (http://cxf.apache.org/) web services framework. We offer both SOAP and REST endpoints. The web interface is built using the Apache Wicket (http://wicket.apache.org/) web application framework. The database is accessed via a Hibernate (http://www.hibernate.org) object-relational mapping layer. The server is running within Sun's Glassfish (http://glassfish.org) application server.

3.9 Future directions

In the near future we would like to extend the interactive facilities of the GPCRDB by offering users more tools to analyze the available data. The entropy-variability analysis pages are a good example of the types of services we will be offering. In addition to our main focus of data collection and integration we would like to extend our focus towards the more challenging field of knowledge integration. The mutation effect predictor is a pilot project to explore the things we can do by combining human expertise with computational power. We are in the process of transforming the GPCRDB from mainly a one-stop resource for GPCRDB data to a place where scientists can use tools to interact with the data and make predictions.

References

1. Overington,J.P., Al-Lazikani,B. and Hopkins,A.L., *How many drug targets are there?* Nat Rev Drug Discov, 2006, **5**: p. 993-6.
2. Vassilatis,D.K., et al., *The G protein-coupled receptor repertoires of human and mouse.* Proc. Natl. Acad. Sci. U.S.A, 2003, **100**: p. 4903-8
3. Klabunde,T. and Hessler,G., *Drug design strategies for targeting G-protein-coupled receptors.* Chembiochem, 2002, **3**: p. 928-44
4. Milligan,G. *G protein-coupled receptor hetero-dimerization: contribution to pharmacology and function.* Br. J. Pharmacol, 2009, **158**: p. 5-14
5. Berman,H.M., et al., *The Protein Data Bank*. Nucleic Acids Res, 2000, **28**: p. 235-42.
6. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res, 2010, **38**, p. D142-48
7. Kanehisa,M., et al., *KEGG for representation and analysis of molecular networks involving diseases and drugs.* Nucleic Acids Res, 2010, **38**: p. D355-60.
8. Stoesser,G., et al., *The EMBL nucleotide sequence database.* Nucleic Acids Res, 2001, **29**: p. 17-21.
9. Benson,D.A., et al., *GenBank.* Nucleic Acids Res, 2008, **36**: p. D25-30
10. Kourist,R., et al., *The alpha/beta-hydrolase fold 3DM database (ABHDB) as a tool for protein engineering.* Chembiochem, 2010, **11**: p. 1635-43
11. Kuipers,R.K., et al. *3DM: systematic analysis of heterogeneous superfamily data to discover protein functionalities.* Proteins, 2010, **78**: p. 2101-13
12. Horn,F., et al., *GPCRDB information system for G protein-coupled receptors*. Nucleic Acids Res, 2003, **31**: p. 294-97.
13. Hull,D., et al., *Taverna: a tool for building and running workflows of services.* Nucleic Acids Res, 2006, **34**: p. W729-32
14. Hekkelman,M.L. and Vriend,G., *MRS: a fast and compact retrieval system for biological data.* Nucleic Acids Res, 2005, **33**: p. W766-769
15. Dowell,R.D., et al., *The distributed annotation system.* BMC Bioinformatics, 2001, 2, 7.
16. Thornton,J. (2009) Annotations for all by all - the BioSapiens network. Genome Biol, 10, 401, 10.1186/gb-2009-10-2-401.
17. Jimenez,R.C., et al., *Dasty2, an Ajax protein DAS client*. Bioinformatics, 2008, **24**: p. 2119-21
18. Brooksbank,C., Cameron,G. and Thornton,J., *The European Bioinformatics Institute's data resources.* Nucleic Acids Res, 2010, **38**: p. D17-25
19. Okuno,Y., et al., *GLIDA: GPCR--ligand database for chemical genomics drug discovery--database and tools update.* Nucleic Acids Res, 2008, **36**: p. D907-912
20. Seeman,P., *Drug Dissociation Constants for Neuroreceptors and Transporters.* Receptor Tables, 1993, **2**.
21. Cutler,D. and Barbier,A., *In brief.* Trends in Pharmacological Sciences, 2002, **23**: p. 258-259.
22. Edvardsen,O., et al., *tGRAP, the G-protein coupled receptors mutant database.* Nucleic Acids Res, 2002, **30**, p. 361-63.
23. Horn,F., Lau,A.L. and Cohen,F.E., *Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors.* Bioinformatics, 2004, **20**: p. 557-68
24. Joosten,R., et al., *PDB_REDO: automated re-refinement of X-ray structure models in the PDB.* Journal of Applied Crystallography, 2009, **42**: p. 376-84.
25. Lohse,M.J. *Dimerization in GPCR mobility and signaling*. Curr Opin Pharmacol, 2010, **10**: p. 53-8
26. Dean,M.K., et al., *Dimerization of G-protein-coupled receptors.* J. Med. Chem, 2001, **44**: p. 4595-614.
27. Hébert,T.E. and Bouvier,M., *Structural and functional aspects of G protein-coupled receptor oligomerization*. Biochem. Cell Biol, 1998, **76**: p. 1-11.
28. Khelashvili,G., et al., *GPCR-OKB: the G Protein Coupled Receptor Oligomer Knowledge Base.* Bioinformatics, 2010, **26**: p. 1804-5
29. Vriend,G., *WHAT IF: a molecular modeling and drug design program.* J Mol Graph, 1990, **8**: p. 52-56
30. Krieger,E., et al., *Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8.* Proteins, 2009, **77** Suppl 9: p. 114-22
31. Oliveira,L., Paiva,A.C.M. and Vriend,G., *Correlated mutation analyses on very large sequence families.* Chembiochem, 2002, **3**: p. 1010-7

32. Oliveira,L., Paiva,A.C., Sander,C. and Vriend,G. *A common step for signal transduction in G protein-coupled receptors.* Trends Pharmacol. Sci, 1994, **15**: p. 170-72.
33. Oliveira,L., Paiva,P.B., Paiva,A.C.M. and Vriend,G., *Sequence analysis reveals how G protein-coupled receptors transduce the signal to the G protein.* Proteins, 2003, **52**: p. 553-60
34. Folkertsma,S., et al., *A family-based approach reveals the function of residues in the nuclear receptor ligand-binding domain.* J. Mol. Biol, 2004, **341**: p. 321-35
35. Oliveira,L., Paiva,P.B., Paiva,A.C.M. and Vriend,G., *Identification of functionally conserved residues with the use of entropy-variability plots.* Proteins, 2003, **52**: p. 544-52
36. Sanders, M.P., et al., *ss-TEA: Entropy based identification of receptor specific ligand binding residues from a multiple sequence alignment of class A GPCRs.* BMC Bioinformatics, 2011. **12**(1): p. 332.
37. Ye,K., Lameijer,E.-W.M., Beukers,M.W. and IJzerman,A.P., *A two-entropies analysis to identify functional positions in the transmembrane region of class A G protein-coupled receptors.* Proteins, 2006, **63**: p. 1018-30
38. L Oliveira, A Paiva and G Vriend (1993) A common motif in G-protein-coupled seven transmembrane helix receptors. J. Comp. Aided Mol. Des., 649-658
39. Cline,M.S., et al., *Integration of biological networks and gene expression data using Cytoscape.* Nat Protoc, 2007, **2**: p. 2366-82
40. Goble,C.A., et al., *myExperiment: a repository and social network for the sharing of bioinformatics workflows.* Nucleic Acids Res, 2010, **38**: p. W677-82
41. Harmar,A.J., et al., *IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels.* Nucleic Acids Res, 2009, **37**: p. D680-85
42. Waterhouse,A.M., et al., *Jalview Version 2--a multiple sequence alignment editor and analysis workbench.* Bioinformatics, 2009, **25**: p. 1189-91
43. Gloriam,D.E., Foord,S.M., Blaney,F.E. and Garland,S.L., *Definition of the G protein-coupled receptor transmembrane bundle binding pocket and calculation of receptor similarities for drug design.* J. Med. Chem, 2009, **52**: p. 4429-42
44. Vroling, B., et al., *Integrating GPCR-specific information with full text articles.* BMC Bioinformatics, 2011, **12**: p. 362
45. Attwood,T.K., et al., *Utopia documents: linking scholarly literature with research data.* Bioinformatics, 2010, **26**: p. i568-i574
46. Braden,M.R., Parrish,J.C., Naylor,J.C. and Nichols,D.E., *Molecular interaction of serotonin 5-HT2A receptor residues Phe339(6.51) and Phe340(6.52) with superpotent N-benzyl phenethylamine agonists.* Mol. Pharmacol, 2006, **70**: p. 1956-64

CHAPTER 4

ss-TEA: entropy based identification of receptor specific ligand binding residues from a multiple sequence alignment of class A GPCRs.

Marijn P. A. Sanders[1], Wilco W.M. Fleuren[1], Stefan Verhoeven[2], Sven van den Beld[1], Wynand Alkema[2], Jacob de Vlieg[1,2] and Jan P. G. Klomp[2,]*

[1]Computational Drug Discovery Group, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands; [2]Department of Molecular Design and Informatics, MSD, Oss, The Netherlands

BMC Bioinformatics, 2011;12; p332

Acknowledgements

The authors thank Sander B. Nabuurs, Peter Groenen and Ross McGuire for critical reading the manuscript and Top Institute Pharma (project number D1-105) for funding.

Abstract

G-protein coupled receptors (GPCRs) are involved in many different physiological processes and their function can be modulated by small molecules which bind in the transmembrane (TM) domain. Because of their structural and sequence conservation, the TM domains are often used in bioinformatics approaches to first create a multiple sequence alignment (MSA) and subsequently identify ligand binding positions. So far methods have been developed to predict the common ligand binding residue positions for class A GPCRs.

Here we present 1) ss-TEA, a method to identify specific ligand binding residue positions for any receptor, predicated on high quality sequence information. 2) The largest MSA of class A non olfactory GPCRs in the public domain consisting of 13324 sequences covering most of the species homologues of the human set of GPCRs. A set of ligand binding residue positions extracted from literature of 10 different receptors shows that our method has the best ligand binding residue prediction for 9 of these 10 receptors compared to another state-of-the-art method.

The combination of the large multi species alignment and the newly introduced residue selection method ss-TEA can be used to rapidly identify subfamily specific ligand binding residues. This approach can aid the design of site directed mutagenesis experiments, explain receptor function and improve modelling. The method is also available online via GPCRDB at http://www.gpcr.org/7tm/ .

4.1 Introduction

G-protein coupled receptors (GPCRs), also known as 7 transmembrane receptors, represent a large superfamily of proteins in the human genome and are responsible for the transduction of an endogenous signal into an intracellular message, which triggers a response in many different physiological pathways. The structural architecture and chemo-mechanical concept of G-protein coupled receptors can be seen as an evolutionarily success as witnessed by the large amount of family members and diversity of applications in biological processes [1].

Not surprisingly, an increasing number of these GPCRs is the subject of investigation as targets in drug discovery. Historical drug discovery approaches have identified GPCRs as a successful drug target, since 25-50% of the drugs currently on the market interact with a GPCR [1, 2].

In humans, the family of 7 transmembrane receptors is represented by approximately 900 members which can be divided in several classes based upon standard similarity searches [3-5].

Recently there has been a reclassification of receptors according to the GRAFS system which has the following groups: glutamate, rhodopsin, adhesion, frizzled/taste2, and secretin[6]. From the structural and functional viewpoint the rhodopsin-like family, also known as the class A receptors, is the largest and best studied family [6].

Receptors from different families are very diverse [1, 5, 7], but can all be characterized by the presence of seven structurally conserved alpha helices, which span the cell membrane. Most GPCRs couple to a G-protein complex upon ligand binding, resulting in the dissociation of the alpha subunit from the beta and gamma subunit. The final signal depends on the alpha subunit of the G-protein (Gαi, Gαs , Gαq/11 , Gα12/13) which is activated and is presumed to be receptor and ligand dependent [8-12]. The non olfactory Class A receptors recognize a large variety of ligands including photons [13], biogenic amines [14], nucleotides [15], peptides [16], proteins [17] and lipid-like substances [18-21]. Most ligands are believed to bind fully or partly within the transmembrane bundle and to trigger signaling through a conserved canonical switch [9]. The assumption that similar molecules bind to similar receptors [22] and that small molecules bind within the upper part of the transmembrane helices, similar to 11-cis retinal in bovine rhodopsin, carazolol in the human beta adrenergic receptor 2, timolol in the turkey beta adrenergic receptor 1 and ZM-241385 in the human adenosine A2 receptor, gives rise to the application of pattern recognition analysis on multiple sequence alignments of those helices or parts thereof to identify ligand binding residues. It has also been shown that for some receptors which bind large proteins, like the luteinizing hormone receptor (LHR), low molecular weight (LMW) compounds can be designed which bind in between the TM-bundle and modify signaling [23, 24], suggesting that the same pattern detection techniques could be used for those receptors as well.

Structure based drug design strategies often rely on high resolution information derived

from protein crystal structures. Elucidating GPCR structures at atomic resolution remains difficult and has only been successful for a small set of receptors so far (bovine rhodopsin [25], squid rhodopsin [26], human beta-2-adrenergic receptor [27], turkey beta-1-adrenergic receptor [28] and the human A2A adenosine receptor [29]). These structures have been extremely helpful for understanding the function and ligand binding properties of class A receptors and are a major step forward towards rational drug design in this class of receptors. However, understanding the differences in for example agonist and antagonist binding or extrapolating structural information on a small subset of GPCRs to evolutionary distant receptors remains problematic and perhaps may only be solved as more structures become available [30]. As long as this information is limited there will be a need for comparative methods to explain the structural and functional differences between GPCRs.

With the recent genome sequencing efforts, more and more data becomes available to perform comparative modelling. Currently, data on 51 species is available in ensemble [31] (release 56) enabling the large scale comparison of sequences within and across species. Methods to mine sequence data and identify structurally and functionally important residues have been developed. For example, in 1996 Lichtarge introduced the evolutionary trace method to calculate the conservation of a residue in each trace of a phylogenetic tree[32]. In 2004 Oliveira et al. introduced the entropy variability plot and showed that the location of the aligned residue positions in these plots correlate to structural characteristics[33]. Based on a similar concept as the entropy variability plot Ye et al. introduced the two entropy analysis (TEA) in 2006 to identify structural and functional positions in the transmembrane region of class A GPCRs[34].

Here we present subfamily specific two entropy analysis (ss-TEA), the first method to identify the ligand binding residues on subfamily level. In contrast to the previously published methods ss-TEA is able to discriminate between subfamilies and able to identify the approximately five residues that are involved in ligand binding for each individual subfamily of the class A GPCRs. ss-TEA is predicated on high quality sequence information deduced from a multiple sequence alignment (MSA) which was generated by extracting species homologues of the class A non olfactory GPCR sequences with a method reported here. This new MSA is characterized by a more complete set of species orthologs which improves the subfamily definition and results of ss-TEA. Receptor specific sets of ligand binding residues, generated by ss-TEA, improve the understanding of receptor ligand interactions and the design of mutagenesis experiments, and guide the process of homology modelling.

4.2 Results and discussion

4.2.1 Sequence retrieval & Alignment

Using a template set of 286 human GPCR sequences, a BLAST search was performed to retrieve non olfactory class A GPCR sequences. This resulted in 20111 sequences originating from 1941 species. An alignment of the transmembrane helices was obtained by gap free alignment of all retrieved sequences using HMM models of the TM domains. Subsequent removal of sequences with low HMM scores resulted in a MSA of 13324 class A GPCR sequences. 33 of the 1941 species contained over 100 class A non olfactory GPCR sequences and were deposited in a database and used for further analysis. The resulting multiple sequence alignment (MSA) comprises 6876 sequences of which 4816 sequences originate from Ensembl and 2060 from Swissprot and TrEMBL. For all aligned helices in the database, it can be shown that the overlap with the predicted helices in Swissprot is over 90% for 90% of the TM sequences and that almost no helices can be found which have less than 75% overlap (data not shown). Due to the gap free alignment procedure of TM domains only those regions are subject to further analysis, loop regions will be omitted and anomalies in helix architecture, i.e. proline induced kinks will not be addressed.

From the distance matrix of all 6876 sequences a hierarchical tree was constructed. A visualization of all human entries of this tree is depicted in **Figure 4.1A**. The number of sequences from all other species, which can be grouped together with a human entry by collapsing a node, is indicated behind the receptor name. Phylogenetic analysis between human and mouse indicated that most human GPCRs have one ortholog in mouse [35]. Since most of the 33 species in our alignment are mammals with an evolutionary distance to human comparable to mouse, it is to be expected that one ortholog from every species can be grouped together with every human receptor. Exceptions to expected 1-1 ortholog pairs will be receptors that have been subject to gene expansion or have become pseudogenes. Examples include the MAS-related G-protein coupled receptors in which gene expansion has occurred, and the GNRHR and 5HT5A receptors which have pseudogenes in human [36]. **Figure 4.1B** shows the distributions of species sequences grouped together in a node with the thyroid stimulating hormone receptor (TSHR). **Figure 4.1C** displays that in most cases one sequence per species is grouped together in a node containing only one human sequence, suggesting that these are orthologs of this human receptor.

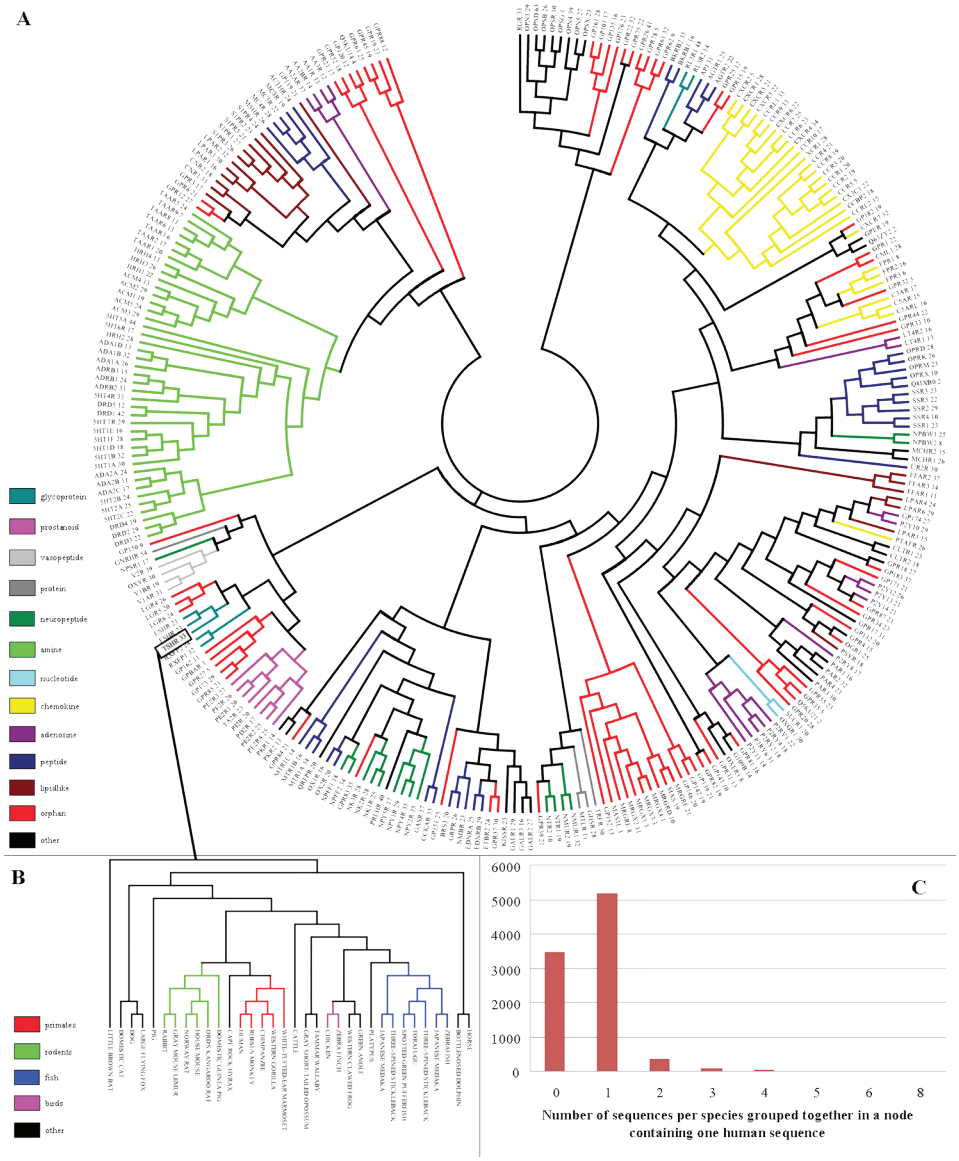


Figure 4.1: Phylogenetic tree of all GPCR sequences. **A:** Visualization of the human entries from the hierarchical tree constructed from the MSA of the TM domains from all sequences in the database. The number indicated after the receptor name equals the number of sequences which are grouped together in the visualized node. Leafs are colored according to the IUPHAR [36] family definition. **B:** Detailed view of the hierarchical tree of the branch including the human thyroid stimulating hormone receptor 1 with the leafs colored according to phylogenetic relatedness. **C:** Distribution of the number of sequences per species grouped together in a node containing one human receptor sequence. The number of missing receptor sequences was calculated with the assumption that each human receptor has one ortholog in each species.

4.2.2. Subfamily definition

To identify ligand binding residues we use a score composed of two entropy values. The underlying hypothesis for this score is that the ligand binding residues are conserved within a subfamily but not across all GPCRs. The residues which are conserved amongst all GPCRs are likely to be structurally important and can be easily identified by a low entropy value for all GPCRs. The size and variability of the subfamily should ensure that apart from structurally important residues only ligand binding residues are conserved within the subfamily. Phylogenetic distance is a measure for the sequence conservation in a subfamily. **Figure 4.2** shows that most of the human receptors in our test set have small phylogenetic distances in subfamilies with sizes towards ~20 sequences. A subfamily of ~20-60 receptors contains homologous receptors (**Figure 4.1**) with slightly larger phylogenetic distances.

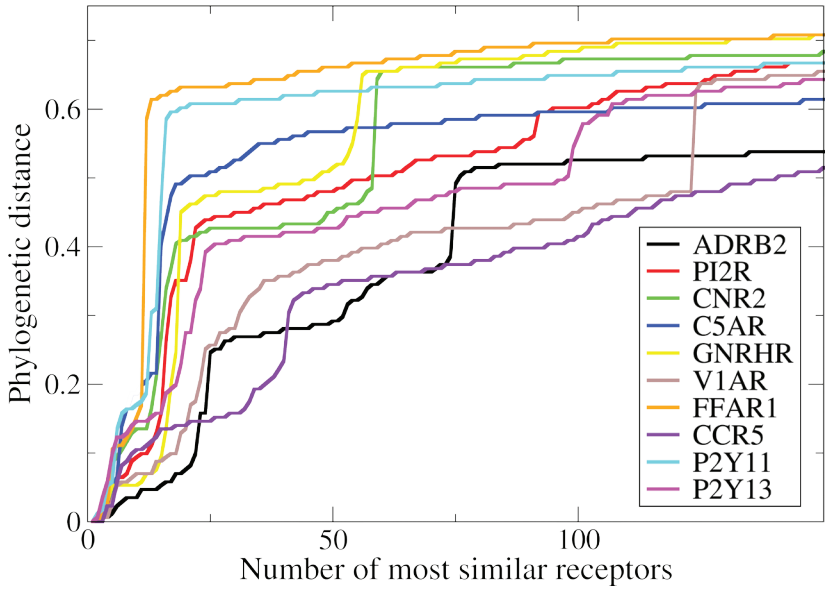


Figure 4.2: Phylogenetic distance towards the human receptor as a function of the xth most similar receptor for the 10 receptors in the test set.

It is impossible to conclude whether or not ligand binding residues are conserved in a subfamily based on solely phylogenetic distances. Important aspects to consider in subfamily selection are that the receptors in a subfamily must bind to relatively similar ligands ensuring evolutionary pressure on the conservation of the residue positions involved in ligand binding, and that evolutionary distances are large enough to observe different amino acid usage amongst residue positions which are not involved in maintaining the structural architecture of the GPCR, signal transduction or ligand binding. We have therefore chosen to calculate the entropy values of all subfamilies with at least 50 and at most 300 sequences.

4.2.3. Reference set

Site directed mutagenesis experiments offer a tool to investigate the function of specific residues in receptors. These experiments have helped to identify residues related to the signal transduction pathway as well as residues involved in ligand binding in GPCRs. Extracting this information from such experiments can however be very complicated, especially if ligands are compared which use different signaling pathways or when agonist are compared to antagonists. Antagonists only have to block active sites and this can be done via interactions with arbitrary residues. Agonists have to trigger certain responses and it is possible that ligands bind to different residues to trigger different responses. Another important aspect in the interpretation of mutation data is to separate direct from indirect effects. Mutations on the membrane facing side of a helix will for example very likely not affect ligand binding in a direct manner, but are more likely to have an influence due to distortion of the secondary structure. We have used site directed mutagenesis data described in literature, to the best of our knowledge, to compile a reference set of ligand binding residues for 10 selected receptors. This reference set consists of 47 residues located at 22 different positions (**Figure 4.3**).

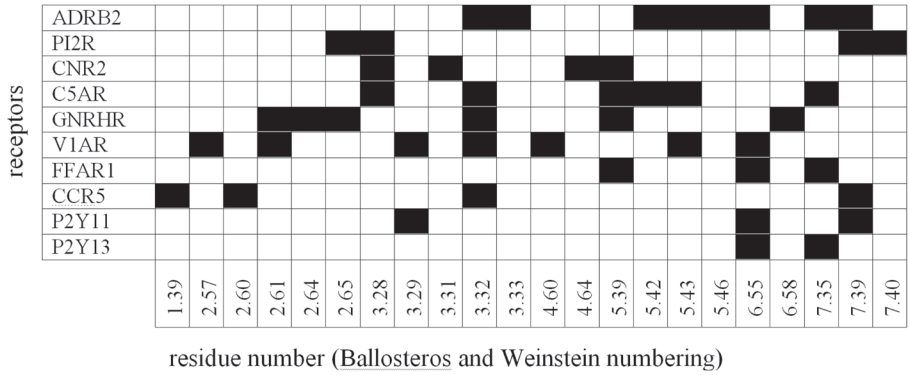


Figure 4.3: Heatmap of reference residues sorted on position and receptor. Ligand binding residues are colored black

4.2.4. Ligand binding residue prediction

Prediction of ligand binding residues as performed by for example evolutionary trace, TEA and Multi-RELIEF is limited to the family level and results in a common description of the structurally important residues and ligand binding pocket. Analyses of the charged aspartate 3.32 in the amine receptors and lysine 7.33 in the opsins, known to be crucial for ligand binding from crystallography, show remarkable conservation characteristics. In both cases the residue is fully conserved inside the family and only rarely observed outside. This suggests that ligand binding residues can be identified by comparing the conservation level of a residue position within a subfamily to the conservation at this

same position for all sequences outside this subfamily. In **Figure 4.4A** the two entropy values reflecting both observations are plotted for the ADRB2 receptor subfamily, which also includes the human receptors ADRB1 and ADRB3. The residues shown to disrupt ligand binding are colored green and are found in the upper left corner as expected. The distance of each residue to this upper left corner is used to rank the residues and used to evaluate the performance of ss-TEA. In **Figure 4.4B** the crystal structure of the ADRB2 receptor, co-crystalized with carazolol (pdbid: 2RH1) is visualized with the residues disrupting ligand binding colored green. The receiver operating characteristic (ROC) curves in **Figures 4.4C** and **4.4D**, plotted with linear and logarithmic x-axis, show the improved ranking of residues according to ligand binding likelihood compared to random ranking. The area under the semi-logarithmic curve (**Figure 4.4D**) was used for further analysis because it puts more emphasis on correctly predicted ligand binding residues in the early phase of the recovery curve. It is typically in this region where performance needs to be outstanding, since many modeling approaches rely heavily on the correct assignment of only a limited number of ligand binding residues.

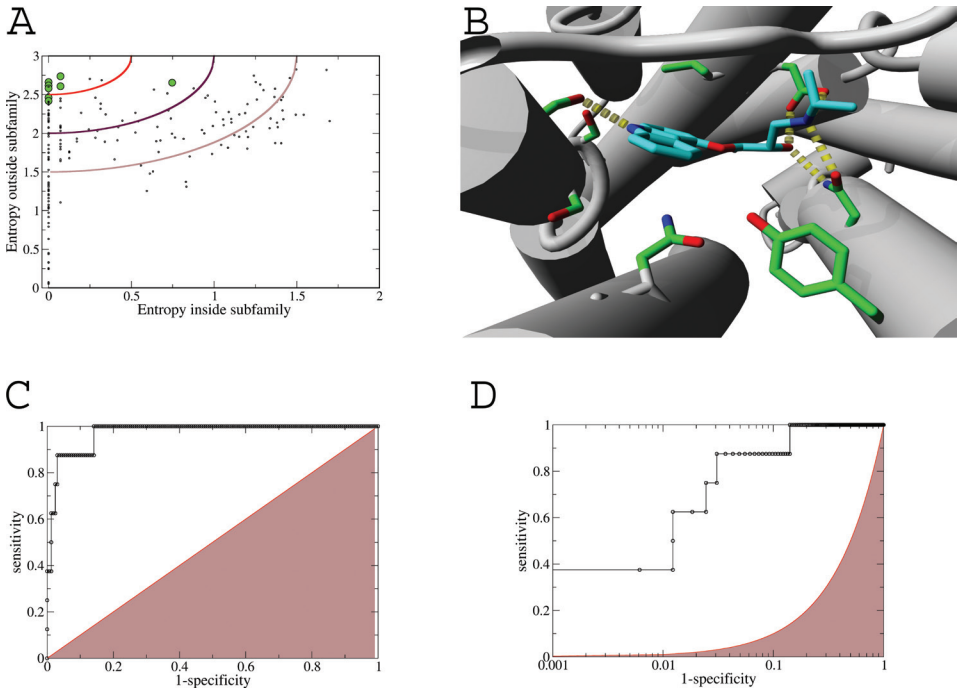


Figure 4.4: Residue selection for the ADRB2 receptor. **A:** Plot of the entropy within the ADRB2 receptor subfamily versus outside the subfamily. Lines are drawn at equal score and residues disrupting ligand binding upon mutation are colored green. **B:** Crystal structure of ADRB2 co-crystalized with carazolol (pdbid: 2RH1), residues disrupting ligand binding upon mutation are colored green. **C:** Receiver Operator Characteristic (ROC) curve showing the ability of ss-TEA to select ligand binding residues compared to random selection. **D:** ROC curve with logarithmic x-axis.

Table 4.1 shows that the mean area under the semi-logarithmic curve of the theoretically optimal ranking and ss-TEA are both 1.9. ss-TEA has the highest score in 7 out of 10 cases compared to the theoretically optimal ranking with the 4 highest scores out of 10 cases. A more realistic example is given by the comparison with the multi-RELIEF + 3d contacts method, which was reported to be the best performing method amongst several state-of-the-art methods [37]. The average pROC AUC of multi-RELIEF is 1.32 if the 22 reference residues are top ranked (see Methods section). ss-TEA gains 0.4 in the pROC AUC compared to multi-Relief in this situation and 0.5 if all residues are taken into account. It is also noteworthy that ss-TEA outperforms multi-Relief for all individual reference receptors except V1AR.

Table 4.1: Area under the semi logarithmic receiver operator curve (pROC AUC) of different rankings of residues for different targets. Top scoring methods are indicated in bold.

Target	Reference set[a]	Multi-Relief	ss-TEA	Theoretically optimal (top-ranked)	Multi-Relief (top-ranked)	ss-TEA (top-ranked)
ADRB2	3.32, 3.33, 5.42, 5.43, 5.46, 6.55, 7.35, 7.39	1.2	1.4	1.9	1.6	1.6
PI2R	2.65, 3.28, 7.39, 7.40	1.0	1.7	1.5	1.4	1.8
CNR2	3.28, 3.31, 4.64, 5.39	0.7	1.5	1.6	1.1	1.8
C5AR	3.28, 3.32, 5.39, 5.42, 5.43, 7.35	1.1	1.5	2.1	1.5	1.7
GNRHR	2.61, 2.64, 2.65, 3.32, 5.39, 6.58	1.4	1.8	1.7	1.7	1.9
V1AR	2.57, 2.61, 3.29, 3.32, 4.60, 5.43, 6.55	1.7	1.3	1.8	1.9	1.5
FFAR1	5.39, 6.55, 7.35	0.9	2.1	2.2	1.3	2.2
CCR5	1.39, 2.60, 3.32, 7.39	1.1	1.5	1.7	1.6	1.9
P2Y11	3.29, 7.39, 6.55	1.5	1.8	1.9	1.8	2.0
P2Y13	6.55, 7.35	1.0	2.1	2.2	1.4	2.2
Average	total 22	1.2	1.7	1.9	1.5	1.9

[a]Ballesteros and Weinstein numbering

Three distinct receptors (ADRB2, CCR5 and GNRHR) which use different residue positions to bind ligands (**Figure 4.3**) have been selected as an example to illustrate the advantage of the subfamily specific approach of ss-TEA.

Figure 4.5 shows in green the residues involved in ligand binding to the ADRB2 receptor only. Likewise, the ligand binding residues for CCR5 and GNRHR are colored blue and red respectively. Residue 7.39 is important for ligand binding in both the ADRB2 and CCR5 receptors and is colored yellow, while position 3.32, colored maroon is a ligand binding residue for all three receptors. **Figure 4.5** shows that green residues are mainly located in the upper left corner of the ADRB2 plot, while the red and blue residues are positioned more to the right. Similar distributions are observed for the blue and red residues in the CCR5 and GNRHR plot respectively, clearly illustrating that the selection of residues by ss-TEA are subfamily specific.

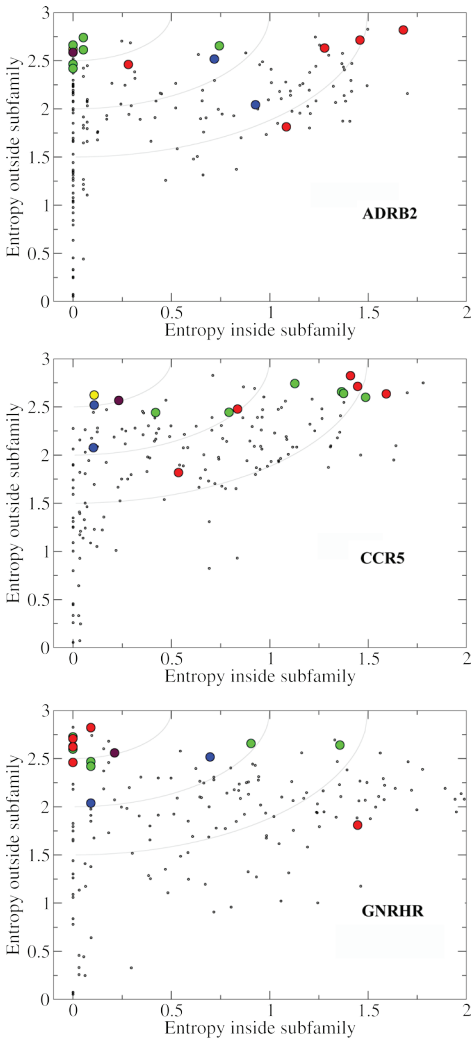


Figure 4.5: ss-TEA plots of ADRB2, CCR5 and GNRHR respectively. Ligand binding residues of the ADRB2 receptor are colored green, CCR5 receptor: blue, GNRHR receptor: red, ADRB2 and CCR5 receptor: yellow and of all three receptors: maroon.

Analyses of the highest ranked residues for all individual human receptors identify subfamily ligand binding characteristics. Determination of the top 10 scoring residues for all human receptors visualized in **Figure 4.6** and colored according to the IUPHAR family definition [36], shows that there is no generic ligand binding mode for class A GPCRs since none of the positions is scored amongst the top 10 for more than 50% of the in total ~300 human receptors. Furthermore it can be seen that that helix I is rarely important for ligand binding, as also observed in the available crystal structures. Even so, some orphan, adenosine and chemokine receptors are characterized by conservation patterns for residues in this helix and might bind ligands with residues from this helix. In addition, the amine receptors can be characterized by the importance of helix three in ligand binding.

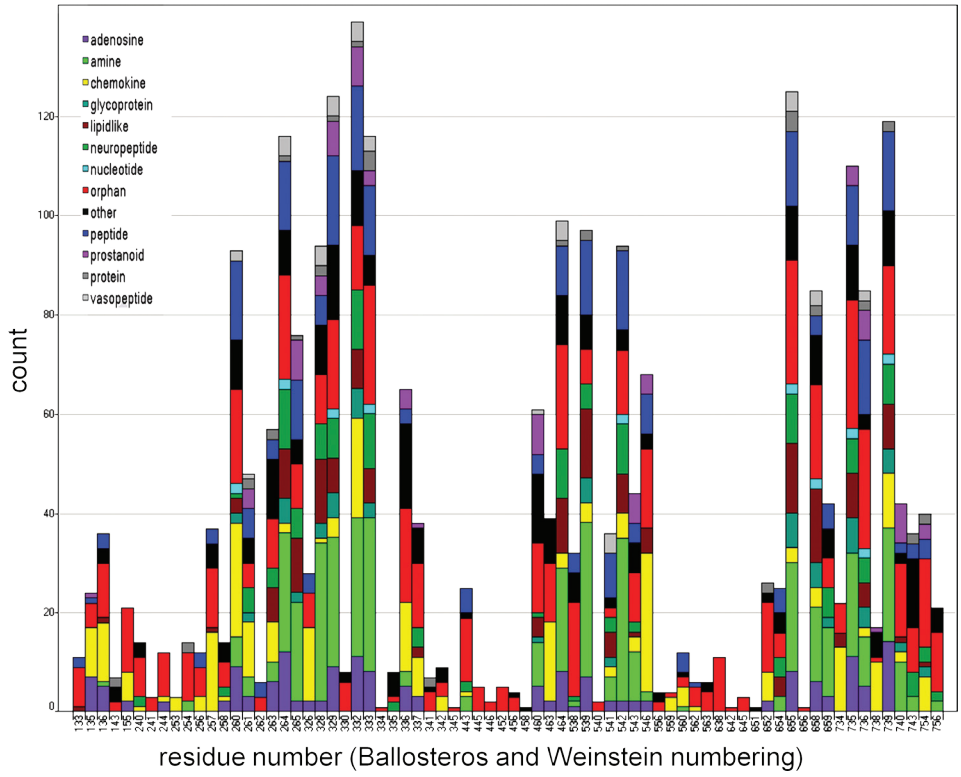


Figure 4.6: Distribution of residue positions scoring amongst the top 10 based on ss-TEA for all human receptors. Bars are colored according to the iuphar family description.

However the comparison of individual receptors within a receptor family also reveals interesting differences in ligand binding behavior. This is illustrated by e.g. position 3.32, which is well conserved in about 50% of all subfamilies, including the aminergic receptors and a subset of the adenosine receptors. For the aminergic receptors it has been proposed that this aspartate is crucial for ligand binding due to its interaction with the positively charged nitrogen of the basic amines, a hypothesis which is confirmed by the crystal structures of ADRB2 and ADRB1. For other receptors this same position is thought to be important for ligand binding involving different amino acids. For example, AA2AR receptor has a conserved valine at position 3.32. Mutation of this valine to alanine or aspartate disrupts ligand binding and illustrates the importance of this conserved valine for this receptor [38]. Position 3.32 ranks at position 45 in the AA1R subfamily, while it ranks at position 11 the AA2AR subfamily suggesting a less important function for the valine in the AA1R receptor, which is indeed confirmed by site directed mutagenesis [38]. Interestingly, receptors with endogenous ligands which completely or largely bind to the N-terminus and/or extracellular loops also demonstrate subfamily specific conservation of residues at the extracellular side of the transmembrane helices. It is remarkable, for example, that 8 of the top 10 ranked residues for the luteinizing hormone receptor are in fact pocket residues. Also noteworthy is that Asp2.64, known to interact with the endogenous ligand [39], is ranked 3rd.

4.3 Conclusions

We have introduced an alignment methodology to create a large multiple sequence alignment of the transmembrane domains of class A non olfactory GPCRs from multiple species. We also introduced a new method to identify ligand binding residues from a MSA, named ss-TEA, and demonstrated the advantage of this new method in combination with the new MSA for the selection of ligand binding residues. The results show the advantage of receptor specific residue selection compared to receptor class specific selection, as well as an improved residue selection for 9 of the 10 reference sets in comparison to the state-of-the-art method Multi-Relief. The large MSA including sequences of multiple species allows us to compare receptors with high sequence similarities and more identical ligand binding profiles which results in a better understanding of the characteristics of those receptors. If more sequence data becomes available for more species, larger alignments can be made, which could possibly even explain differences between close homologs. Our alignment in combination with the residue selection method described here can be used to quickly identify ligand binding residues. This can subsequently be used to design site directed mutagenesis experiments, explain receptor function and improve modelling. The ss-TEA predictions for class A GPCRs can be accessed via GPCRDB at www.gpcr.org/7tm/.

4.4 Methods

Our approach makes use of different input sources which are connected via algorithms as outlined in **Figure 4.7**. All steps will be outlined and discussed in sequential order below.

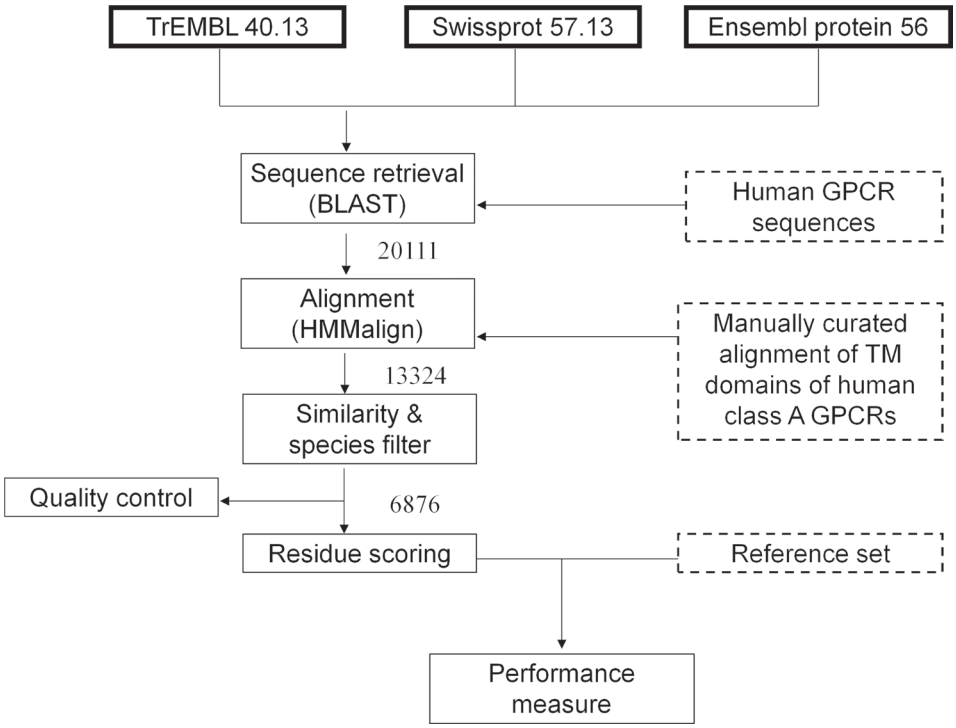


Figure 4.7: Schematic flowchart of the methodology to create the alignment, score the residues and evaluate the performance of the residue selection method. Publicly available data sources are indicated with a bold border style, in-house data with dashed a border style and methods with a normal border style. The numbers indicate the number of sequences which is used at each step.

4.4.1. Sequence retrieval

The first step in our approach is to extract GPCR sequences for different species from available data sources. To obtain sequences we performed a BLAST [40] search with 286 manually curated query sequences from human class A non-olfactory GPCRs against Swissprot, Ensembl and TrEMBL. All query sequences were blasted against Swissprot 57.13 [41, 42], Translated EMBL (TrEMBL) 40.13 [42, 43] and Ensembl Protein 56 [31], using the BLOSUM62 scoring matrix, an expected cutoff of 10 and word size 3. Furthermore, a gap opening penalty of 11 and a gap extension penalty of 1 were used. Finally, we selected all sequences with an e-value < 0.01, subject length identity > 25%, alignment identity > 40% and a minimal query length of 20 amino acids.

4.4.2. Alignment

The available GPCR crystal structures have shown that all helices can be structurally aligned without introducing gaps in the sequence alignment. For this reason a manually curated gap-free alignment of the TM domains of the human class A non-olfactory GPCR sequences was created and used to construct a Hidden Markov Model (HMM) for each separate helix, using HMMbuild (HMMER [44] 2.3.2 (Oct 2003)) with default settings. Subsequently each hmm model of each helix was aligned against all extracted sequences from the previous step, without allowing the introduction of gaps, using HMMalign. Alignments which had an incorrect helix ordering were subsequently extracted and subject to realignment on a smaller part of the sequence. A typical example is the realignment of one helix on the sequence in between two correctly aligned neighbor helices (**Figure 4.8**). As a final filter, all sequences with a low similarity score to the hmm model for over 4 out of 7 helices were discarded. A threshold of 4 was chosen, since a few annotated human sequences, e.g., the prostanoids, were shown to have weak patterns for up to 4 helices. The threshold for the similarity score of each individual helix was set after the compilation of artificial sequences with an identical amino acid distribution for each helix as in the manually curated alignment of human GPCRs. These artificial sequences were subsequently aligned to the previously built hmm model of the helix, and the threshold for the helix was set to the score at which 95% of the artificial sequences fails to pass. Low sequence quality may cause duplicate entries of the same receptor and species. To avoid these duplicates, all but 1 sequence, of all sets of sequences of an individual species which had less than 10 amino acids difference, were removed.

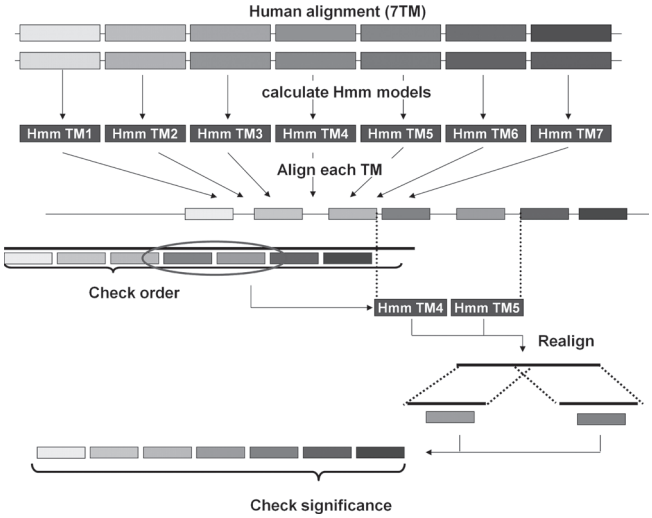


Figure 4.8: Schematic representation of the alignment procedure. First a HMM model is calculated for each TM domain from a manually curated alignment of the 7TM domains of 286 human GPCRs. Each HMM model is subsequently aligned to each GPCR sequence after which the ordering of the aligned helices is checked. In case of an incorrect ordering realignment is performed on smaller parts of the sequence. Finally the significance of each aligned helix is checked.

4.4.3. Database

Incomplete sequencing of the genomes of many species causes bias towards certain receptor subfamilies. To prohibit such bias, all sequences of species with less than 100 amino acid sequences of GPCRs were removed from the MSA. All GPCR sequences of species of which at least 100 different sequences were obtained, were stored in a database and used in all analysis discussed below. To enable querying on a higher level than the individual sequences, a hierarchical tree of the phylogenetic distance matrix calculated from the alignment of all 7 TMs of all receptors was created, using the neighbor joining algorithm as implemented in clustalW [45] 2.0.11 with a 100 fold bootstrap. The sequences which group together at a node in this tree, a so called subfamily, can be queried for their properties.

4.4.4. Residue selection

To perform knowledge based residue selections which reflect the likelihood of residues being involved in ligand binding, we added two Shannon entropy scores for each alignment position of each receptor to the database. One entropy value reflects the conservation of a position inside the subfamily (E^{in}) while the other entropy reflects the conservation of this same position in all sequences which do not belong to this subfamily (E^{out}). The Shannon entropy itself is given by:

$$E_i = -\sum_{a=1}^{20} F_{ia} \ln F_{ia} \qquad (1)$$

With

$$F_{ia} = Number_{ia} / m \qquad (2)$$

$Number_{ia}$ is the number of sequences with residue type *a* at alignment position *i*. Others have already suggested that ligand binding residues can be obtained from both calculated entropy values [33, 34]. Therefore we introduce one score which combines both calculated entropies.

$$S_i = \sqrt{\left(E_i^{in}\right)^2 + \left(\ln(20) - E_i^{out}\right)^2} \qquad (3)$$

A final score for each residue position was calculated after evaluation of the score at multiple branches of the hierarchical tree using:

$$F_i = \min(S_i, cuttree(j)) \ \text{ for all j} \in [50,300] \qquad (4)$$

where *j* reflects the number of sequences selected in the branch. To validate the performance we finally ranked all residues according to the score with the minimum scoring residue at rank 1.

4.4.5. Reference Set

Site directed mutagenesis data is available for many GPCRs with different levels of detail depending on the research question. In this paper ten well studied and evolutionary diverse Class A GPCRs are used for which extensive site directed mutagenesis data exists as well as a binding model based on these data. For each of the receptors a reference set of residues crucial for ligand binding was compiled using the mutation data described in GPCRdb [5] and literature models of the binding mode. The choice of receptors from different branches of the sequence tree was made to emphasize the advantage of a method able to identify different ligand binding residues for different receptors and to show that the method does not have a bias towards certain subfamilies. The receptors in the reference set are; beta-2 adrenergic receptor (ADRB2) [27, 46]; Prostacyclin receptor (PI2R) [47]; C5a anaphylatoxin chemotactic receptor (C5AR) [48]; Cannabinoid receptor 2 (CNR2) [49, 50]; Gonadotropin-releasing hormone receptor (GNRHR) [51]; Vasopressin V1a receptor (V1AR) [24]; Free fatty acid receptor1 (FFAR1) [52]; C-C Chemokine receptor type 5 (CCR5) [53]; P2Y purinoceptor 11 [54] and 13 [55] (P2Y11, P2Y13). Residues that were not part of the pocket [56] were neglected as well as mutations which are debatable because of different effects using different ligands or because results were not consistent in different measurements. The final selection only includes residues with substantial effect on ligand binding. The A2A adenosine receptor was deliberately not used as a reference set in this study, since site directed mutagenesis data and the crystal structure suggest that there is no general, family conserved receptor binding pocket for the A2A adenosine receptor [29, 38].

4.4.6. Performance measure (Area Under the Log Curve)

The performance of our residue ranking method is assessed using the Area under the semi-logarithmic receiver operating characteristic (ROC) curve [57]. This method favors true ligand binding residues early in the recovery curve and is calculated using:

$$pROC\ AUC = \frac{1}{n}\sum_{i}^{n}\log_{10}\left(\frac{1}{\beta_i}\right) \qquad (5)$$

Where n is the number of true ligand binding residues and β_i is the false positive frequency corresponding to the point at which the ith true residue is found. β_i is typically calculated as the fraction of false positives which is ranked higher than the ith true positive. The score of the pROC AUC corresponding to a random selection is 0.434 and is unbounded on the high side. A perfect ordering of ligand binding residues amongst 100 non ligand binding residues will for example score 2.0.

4.4.7. Benchmark

To illustrate the advantage of subfamily specific ranking over generic ranking we compiled a theoretically optimal generic ranking of ligand binding residues. This ranking is created by ordering the residues of ten different receptors according to the number of receptors which use these positions for ligand binding. The ranking of positions used by the same number of receptors is arbitrary, potentially altering the results, although it is expected to have only a minor effect. Because the theoretically compiled optimal ranking includes information about the location of the pocket we also included this information in the ss-TEA and Multi-Relief method and scored the 22 residues included in the theoretically compiled optimal ranking prior to all other residues. The rankings which include this information will be indicated in this paper as top ranked. As a benchmark we compared our top ranking to both the theoretically compiled optimal ranking and Multi-Relief + 3d contacts top ranking [37]. Briefly, Multi-Relief takes a multiple sequence alignment and predefined subfamily ontology as input, then iteratively selects 2 subfamilies and optimizes a weight vector able to optimally separate the sequences from both [37]. The optimization of a single weight vector in the iterative process results in one vector able to discrimate between all provided classes. The weight of a residue in the Multi-Relief + 3d contacts method can be altered towards its local environment as obtained from recent crystal structures.

References

1. Klabunde, T. and G. Hessler, *Drug design strategies for targeting G-protein-coupled receptors.* ChemBioChem, 2002. **3**(10): p. 928-44.
2. Hopkins, A.L. and C.R. Groom, *The druggable genome.* Nat Rev Drug Discov, 2002. **1**(9): p. 727-30.
3. Attwood, T.K. and J.B. Findlay, *Fingerprinting G-protein-coupled receptors.* Protein Eng, 1994. **7**(2): p. 195-203.
4. Kolakowski, L.F., Jr., *GCRDb: a G-protein-coupled receptor database.* Receptors Channels, 1994. **2**(1): p. 1-7.
5. Horn, F., et al., *GPCRDB information system for G protein-coupled receptors.* Nucleic Acids Res, 2003. **31**(1): p. 294-7.
6. Jacoby, E., et al., *The 7 TM G-protein-coupled receptor target family.* ChemMedChem, 2006. **1**(8): p. 761-82.
7. Foord, S.M., et al., *International Union of Pharmacology. XLVI. G protein-coupled receptor list.* Pharmacol Rev, 2005. **57**(2): p. 279-88.
8. Kenakin, T., *Efficacy at G-protein-coupled receptors.* Nat Rev Drug Discov, 2002. **1**(2): p. 103-10.
9. Christopoulos, A., *Allosteric binding sites on cell-surface receptors: novel targets for drug discovery.* Nat Rev Drug Discov, 2002. **1**(3): p. 198-210.
10. Perez, D.M. and S.S. Karnik, *Multiple signaling states of G-protein-coupled receptors.* Pharmacol Rev, 2005. **57**(2): p. 147-61.
11. Maudsley, S., B. Martin, and L.M. Luttrell, *The origins of diversity and specificity in g protein-coupled receptor signaling.* J Pharmacol Exp Ther, 2005. **314**(2): p. 485-94.
12. Urban, J.D., et al., *Functional selectivity and classical concepts of quantitative pharmacology.* J Pharmacol Exp Ther, 2007. **320**(1): p. 1-13.
13. Okada, T., et al., *Activation of rhodopsin: new insights from structural and biochemical studies.* Trends Biochem Sci, 2001. **26**(5): p. 318-24.
14. Vernier, P., et al., *An evolutionary view of drug-receptor interaction: the bioamine receptor family.* Trends Pharmacol Sci, 1995. **16**(11): p. 375-81.
15. Fredholm, B.B., et al., *International Union of Pharmacology. XXV. Nomenclature and classification of adenosine receptors.* Pharmacol Rev, 2001. **53**(4): p. 527-52.
16. Janecka, A., J. Fichna, and T. Janecki, *Opioid receptors and their ligands.* Curr Top Med Chem, 2004. **4**(1): p. 1-17.
17. Horuk, R., *Chemokine receptors.* Cytokine Growth Factor Rev, 2001. **12**(4): p. 313-35.
18. Brown, A.J., S. Jupe, and C.P. Briscoe, *A family of fatty acid binding receptors.* DNA Cell Biol, 2005. **24**(1): p. 54-61.
19. Chun, J., et al., *International Union of Pharmacology. XXXIV. Lysophospholipid receptor nomenclature.* Pharmacol Rev, 2002. **54**(2): p. 265-9.
20. Brink, C., et al., *International Union of Pharmacology XLIV. Nomenclature for the oxoeicosanoid receptor.* Pharmacol Rev, 2004. **56**(1): p. 149-57.
21. Kostenis, E., *A glance at G-protein-coupled receptors for lipid mediators: a growing receptor family with remarkably diverse ligands.* Pharmacol Ther, 2004. **102**(3): p. 243-57.
22. Klabunde, T., *Chemogenomic approaches to drug discovery: similar receptors bind similar ligands.* Br J Pharmacol, 2007. **152**(1): p. 5-7.
23. van Koppen, C.J., et al., *A signaling-selective, nanomolar potent allosteric low molecular weight agonist for the human luteinizing hormone receptor.* Naunyn Schmiedebergs Arch Pharmacol, 2008. **378**(5): p. 503-14.
24. Mouillac, B., et al., *The binding site of neuropeptide vasopressin V1a receptor. Evidence for a major localization within transmembrane regions.* J Biol Chem, 1995. **270**(43): p. 25771-7.
25. Palczewski, K., et al., *Crystal structure of rhodopsin: A G protein-coupled receptor.* Science, 2000. **289**(5480): p. 739-45.
26. Murakami, M. and T. Kouyama, *Crystal structure of squid rhodopsin.* Nature, 2008. **453**(7193): p. 363-7.
27. Cherezov, V., et al., *High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor.* Science, 2007. **318**(5854): p. 1258-65.
28. Warne, T., et al., *Structure of a beta1-adrenergic G-protein-coupled receptor.* Nature, 2008. **454**(7203): p. 486-91.
29. Jaakola, V.P., et al., *The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist.* Science, 2008. **322**(5905): p. 1211-7.

30. Cavasotto, C.N. and S.S. Phatak, *Homology modeling in drug discovery: current trends and applications.* Drug Discov Today, 2009. **14**(13-14): p. 676-83.
31. Hubbard, T.J., et al., *Ensembl 2009.* Nucleic Acids Res, 2009. **37**(Database issue): p. D690-7.
32. Madabushi, S., et al., *Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions.* J Biol Chem, 2004. **279**(9): p. 8126-8132.
33. Oliveira, L., et al., *Sequence analysis reveals how G protein-coupled receptors transduce the signal to the G protein.* Proteins, 2003. **52**(4): p. 553-60.
34. Ye, K., et al., *A two-entropies analysis to identify functional positions in the transmembrane region of class A G protein-coupled receptors.* Proteins, 2006. **63**(4): p. 1018-30.
35. Bjarnadottir, T.K., et al., *Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse.* Genomics, 2006. 88(3): p. 263-73.
36. Harmar, A.J., et al., *IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels.* Nucleic Acids Res, 2009. **37**(Database issue): p. D680-5.
37. Ye, K., et al., *Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting.* Bioinformatics, 2008. **24**(1): p. 18-25.
38. Kim, S.K., et al., *Modeling the adenosine receptors: comparison of the binding domains of A2A agonists and antagonists.* J Med Chem, 2003. **46**(23): p. 4847-59.
39. Ji, I., H. Zeng, and T.H. Ji, *Receptor activation of and signal generation by the lutropin/choriogonadotropin receptor. Cooperation of Asp397 of the receptor and alpha Lys91 of the hormone.* J Biol Chem, 1993. **268(**31): p. 22971-4.
40. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
41. *The Universal Protein Resource (UniProt) in 2010.* Nucleic Acids Res, 2010. **38**(Database issue): p. D142-8.
42. Jain, E., et al., *Infrastructure for the life sciences: design and implementation of the UniProt website.* BMC Bioinformatics, 2009. **10**: p. 136.
43. Ballesteros, J.A.W., H., *Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein coupled receptors.* Methods Neurosci., 1995. **25**: p. 366-428.
44. Eddy, S.R. *HMMER: Profile hidden Markov models for biological sequence analysis*. Available from: http://hmmer.janelia.org.
45. Larkin, M.A., et al., *Clustal W and Clustal X version 2.0.* Bioinformatics, 2007. **23**(21): p. 2947-8.
46. Rosenbaum, D.M., et al., *GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function.* Science, 2007. **318**(5854): p. 1266-73.
47. Stitham, J., et al., *The unique ligand-binding pocket for the human prostacyclin receptor. Site-directed mutagenesis and molecular modeling.* J Biol Chem, 2003. **278**(6): p. 4250-7.
48. Gerber, B.O., et al., *An activation switch in the ligand binding pocket of the C5a receptor.* J Biol Chem, 2001. **276**(5): p. 3394-400.
49. Poso, A. and J.W. Huffman, *Targeting the cannabinoid CB2 receptor: modelling and structural determinants of CB2 selective ligands.* Br J Pharmacol, 2008. **153**(2): p. 335-46.
50. Raitio, K.H., et al., *Targeting the cannabinoid CB2 receptor: mutations, modeling and development of CB2 selective ligands.* Curr Med Chem, 2005. 12(10): p. 1217-37.
51. Millar, R.P., et al., *Gonadotropin-releasing hormone receptors.* Endocr Rev, 2004. **25**(2): p. 235-75.
52. Sum, C.S., et al., *Identification of residues important for agonist recognition and activation in GPR40.* J Biol Chem, 2007. **282**(40): p. 29248-55.
53. Paterlini, M.G., *Structure modeling of the chemokine receptor CCR5: implications for ligand binding and selectivity.* Biophys J, 2002. **83**(6): p. 3012-31.
54. Costanzi, S., et al., *Architecture of P2Y nucleotide receptors: structural comparison based on sequence analysis, mutagenesis, and homology modeling.* J Med Chem, 2004. **47**(22): p. 5393-404.
55. Ivanov, A.A., S. Costanzi, and K.A. Jacobson, *Defining the nucleotide binding sites of P2Y receptors using rhodopsin-based homology modeling.* J Comput Aided Mol Des, 2006. **20**(7-8): p. 417-26.
56. Gloriam, D.E., et al., *Definition of the G protein-coupled receptor transmembrane bundle binding pocket and calculation of receptor similarities for drug design.* J Med Chem, 2009. **52**(14): p. 4429-42.
57. Clark, R.D. and D.J. Webster-Clark, *Managing bias in ROC curves.* J Comput Aided Mol Des, 2008. **22**(3-4): p. 141-6.

CHAPTER 5

Snooker: a structure-based pharmacophore generation tool applied to class A GPCRs

Marijn P. A. Sanders[1], Stefan Verhoeven[2], Chris de Graaf[3], Luc Roumen[3], Bas Vroling[4], Sander B. Nabuurs[1], Jacob de Vlieg[1,2] and Jan P.G. Klomp[2]

[1]Computational Drug Discovery Group, CMBI, Radboud University Nijmegen, Nijmegen, The Netherlands; [2]Department of Molecular Design and Informatics, MRL, MSD, Oss, The Netherlands; [3]Division of Medicinal Chemistry, LACDR, VU University Amsterdam, Amsterdam, The Netherlands; [4]Modeling and Data Mining Group, CMBI, Radboud University Nijmegen, Nijmegen, The Netherlands

Journal of Chemical Information and Modeling, 2011; 51; 2277-2292

Acknowledgements

The authors thank Peter Groenen, Ross McGuire, Scott Lusher and Andreas Bender for critical reading the manuscript. This work was performed within the framework of Dutch Top Institute Pharma, project “The GPCR Forum: Novel concepts and tools for established targets (project nr. D1-105)”.

Abstract

G-protein coupled receptors (GPCRs) are important drug targets for various diseases and of major interest to pharmaceutical companies. The function of individual members of this protein family can be modulated by the binding of small molecules at the extracellular side of the structurally conserved transmembrane (TM) domain. Here, we present Snooker, a structure-based approach to generate pharmacophore hypotheses for compounds binding to this extracellular side of the TM domain. Snooker does not require knowledge of ligands, is therefore suitable for apo-proteins and can be applied to all receptors of the GPCR protein family. The method comprises the construction of a homology model of the TM domains and prioritization of residues on the probability of being ligand binding. Subsequently, protein properties are converted to ligand space and pharmacophore features are generated at positions where protein ligand interactions are likely. Using this semi-automated knowledge-driven bioinformatics approach we have created pharmacophore hypotheses for 15 different GPCRs from several different subfamilies. For the beta-2-adrenergic receptor we show that ligand poses predicted by Snooker pharmacophore hypotheses reproduce literature supported binding modes for ~75% of compounds fulfilling pharmacophore constraints. All 15 pharmacophore hypotheses represent interactions with essential residues for ligand binding as observed in mutagenesis experiments and compound selections based on these hypotheses are shown to be target specific. For 8 out of 15 targets enrichment factors above 10 fold are observed in the top 0.5% ranked compounds in a virtual screen. Additionally, prospectively predicted ligand binding poses in the human dopamine D3 receptor based on Snooker pharmacophores were ranked amongst the best models in the community wide GPCR dock 2010 assessment.

5.1. Introduction

G-protein coupled receptors (GPCRs) represent a large superfamily of membrane proteins responsible for the signal transduction from the extracellular to intracellular side of the cell membrane in many different physiological pathways. Therefore, they are effective drug targets for various diseases and of major interest to pharmaceutical companies [1, 2].

The GPCR family is characterized by seven conserved alpha-helices, which span the cell membrane. To date the crystal structure of only six different GPCRs have been elucidated because proteins from this family are, like many other membrane proteins, difficult to crystallize. The lack of high resolution structural data complicates the process of rational drug design especially for large and structurally diverse families such as the GPCRs. In modern drug discovery, computer-aided techniques are often used to speed up the design process. These techniques are typically divided into ligand-based and structure-based approaches. Ligand-based drug design techniques rely on the availability of known active compounds and have proven to be very successful in the design of new compounds [3-8]. Commonly used ligand-based techniques that are frequently combined with structure-based approaches include the use of a spatial arrangement of key chemical features, a so called pharmacophore, to discriminate active from inactive compounds. A disadvantage of ligand derived pharmacophore hypotheses is the assumption that active compounds bind in a similar binding mode, and consequently, the designed compounds usually have less novelty [9].

Structure-based drug design, on the other hand, does require a three-dimensional structure of the protein, acquired by means of several techniques, including electron microscopy, atomic force microscopy, X-ray crystallography and NMR spectroscopy, or by computational methods e.g. homology modeling. Depending on the accuracy of a three-dimensional model, structure-based searching strategies, such as docking, have proven to be very successful in the design of new active compounds [10-13] also for GPCRs [14-16]. However, working with less accurate structures, as typically obtained by homology modeling, remains a major challenge. New developments, such as induced fit docking, have increased the accuracy of results [17], but are computationally expensive and remain dependent to some extent on prior knowledge.

A protein-based approach that depends less on protein structure resolution is the association of sequence motifs with ligand (interaction) features. In the thematic analysis method [18], structure activity relationships (SAR) of class A and B GPCRs were used to generate a pairing of sequence patterns/themes and ligand structural motifs. Next, focused libraries can be designed by inclusion of compounds with structural motifs which occur in ligands for receptors which share similar sequence patterns. This method contains limited three dimensional structural information and is limited to the structural motifs observed in known ligands.

Another low resolution approach in structure-based drug design is the use of

pharmacophore models derived from protein binding sites. These structure based pharmacophore models can be derived from homology models and have been successfully applied for characterizing ligand binding pockets and virtual screening for ligands fort various protein targets, including GPCRs [14, 19-22]. Kratochwil et al. [23] developed a method to characterize a GPCR pocket with 35 pharmacophore features representing the 35 residues aligning the ligand binding pocket.

Klabunde et al. presented an approach to construct structure-based pharmacophore hypotheses for class A GPCRs based on chemoprints of the pharmacophore hypotheses derived from 10 different homology models and 3 X-ray structures of GPCRs [24].

Integrating ligand- and structure-based approaches allows for optimal use of all available data and has resulted in a number of successful virtual screens in which new compounds have been identified [25, 26].

Recently, the structures of several new GPCRs together with an increasing amount of annotated ligand data have been available in databases such as PDB, CheMBL [27], DrugBank [28], BindingDB [29-31], PDBBind [32, 33], MOAD [34, 35], WOMBAT [36] and Glida-DB [37]. All these data repositories allow data mining to construct for example structure activity relationships (SAR), including Glida-DB and GPCR SARfari (based on CheMBL), which are dedicated to GPCRs. However, they generally do not provide structural understanding of receptor-ligand interactions and offer only limited tools for the discovery of new chemical entities. There is an urging need for tools to connect structure and ligand data especially in the field of GPCRs where only limited crystal structures are available.

We present here Snooker, a low resolution approach whereby structure-based pharmacophore hypotheses are constructed with no prior knowledge of ligand structure or interactions. The pharmacophore hypotheses are obtained from homology models constructed on-the-fly, based on an in-house sequence alignment of the seven transmembrane domains [38] and a crystal structure template. Residues important for ligand binding are identified by analysis of Shannon entropies of structurally conserved positions in a multiple sequence alignment and chemical features representing protein-ligand interactions are positioned in the binding pocket. This abstraction of protein-ligand interaction properties into pharmacophores allows for the discovery of new active compounds and connects structural knowledge to ligand data.

The validity of our pharmacophores is tested in three different experiments. First of all, poses of known active compounds are reproduced in a retrospective ligand binding mode prediction experiment in the beta-2 adrenergic receptor (ADRB2). Secondly, the eticlopride binding mode in the dopamine D3 receptor (DRD3) is successfully predicted in the community wide GPCR DOCK 2010 assessment prior to the release of the DRD3 co-crystal structure. Finally, structure-based pharmacophore models of 15 different GPCRs (matching key ligand binding residues) are successfully used to identify target specific ligand sets and discriminate active from inactive compounds in a virtual cross-screen.

5.2. Outline of the approach

The Snooker approach consists of several stages outlined and discussed in sequential order below. **Figure 5.1** shows the construction of a pharmacophore model and an example of a subsequent pharmacophore search. For clarity and reference, **Figure 5.1** shows the results obtained at each different stage of pharmacophore construction for the human beta2 adrenoceptor based on the rhodopsin crystal structure co-crystalized with retinal (Protein Data Bank (PDB) entry 1GZM [39]) and the subsequent pharmacophore search of R-R-formoterol using the obtained pharmacophore.

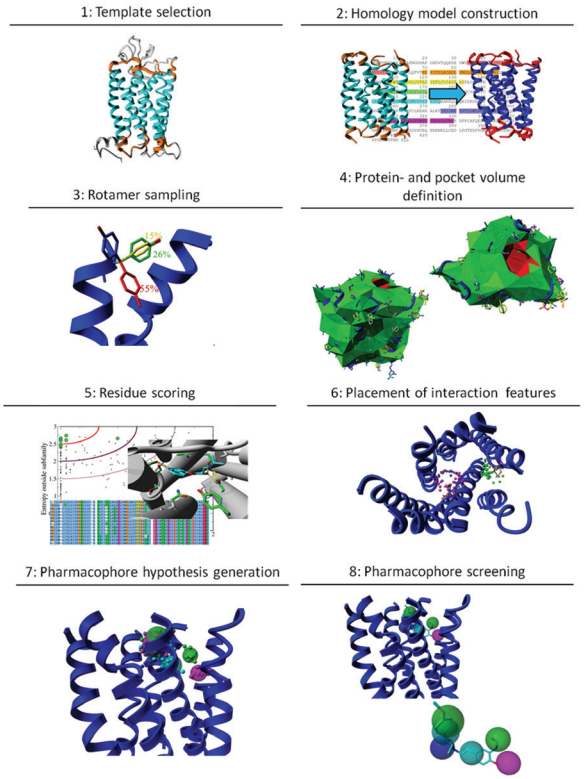


Figure 5.1: Visual outline of the Snooker approach, illustrated using the pharmacophore hypothesis generation for the human beta 2 adrenergic receptor based on a bovine rhodopsin crystal structure (pdb code: 1GZM) and the subsequent positioning of R-R-formoterol in the pharmacophore. Starting with the crystal structure of bovine rhodopsin, (1) the structurally conserved alpha-helices (cyan) and five residues at each side of each helix (orange) are extracted. A homology model (2) is constructed based on the alignment of the model receptor sequence with this template. An alpha helix specific rotamer library is used to add a rotamer ensemble (3). The pocket is detected (4) by a Delaunay tessellation of the Cα-atom and average side chain atom positions. Residues are scored upon ligand binding probability (5) by multiple sequence alignment analysis, and 'interaction' points (6) are placed inside the pocket volume using in literature described interaction geometries with densities corresponding to the residue score and rotamer probability. Next, pharmacophore features are generated (7) with a fuzzy pharmacophore algorithm applied on the interaction points. Finally, (8) ligands fulfilling all pharmacophore constraints are aligned to the pharmacophore. In this example, R-R-formoterol matches a pharmacophore comprised of 6 features.

5.2.1. Template selection

The initial stage of the Snooker protocol starts with the selection of one or multiple template structures. When this research was conducted only the structures of the rhodopsin, adenosine A2A and the beta 1 and 2 adrenergic receptor were available. However future crystal structures and custom made homology models or combinations thereof can be used as template structures as well. Because loop modeling for GPCRs remains very challenging, the templates are stripped down until only the 7 transmembrane helices remain. Inclusion of loops or even single residues is still possible but requires manual adjustments of the provided templates and is only recommended for highly similar loop regions or where prior knowledge on the target is available. The orange and cyan colored residues of the template shown in panel 1 of **Figure 5.1** are used as the starting point for the construction of the homology model.

5.2.2. Homology model construction

The 7 transmembrane (TM) domain sequences of the desired GPCR are extracted from an in-house gap-free multiple sequence alignment of those 7TM domains [38] and is used together with the template sequence and structure to construct the homology model. The backbone and conserved residues are kept rigid during this stage. Placement of non-conserved residues is based upon a position-specific rotamer library [41]. Finally the hydrogen bond network is optimized and bumps are removed.

5.2.3. Rotamer sampling.

For all residues we include the most likely rotamers from an alpha-helix specific rotamer library [42] to account for possible model inaccuracy of the initial homology model. During the generation of these new rotameric states, clashes are allowed, with the intent to maximize sampling. The procedure stores the probability of each rotamer in the ensemble for later use in the protein- and pocket volume definition as well as in the placement of the interaction features. The use of an ensemble of rotamers avoids the computational magnitude that would result from considering all possible models which can be obtained by combining all possible rotamer states for each single residue. The rotamer ensemble is subsequently used to generate both the protein- and pocket volume definition.

5.2.4. Protein- and pocket volume definition.

To restrict the placement of pharmacophoric features to the ligand binding pocket, both pocket and protein volumes are constructed. Traditional pocket detection techniques like grid-based sampling methods require high resolution structures, cannot deal with ensemble structures and are extremely sensitive to small deviations, and are therefore less suitable for our approach. Thus, we have chosen to describe each rotamer ensemble with the position of the Cα-atom and an estimated point representing the sidechain.

This estimated point is effectively an averaging over all rotamers which will describe the sidechain almost exactly if just one rotamer exists, and will approximate the Cα-atom for a diverse rotamer ensemble.

5.2.5. Residue scoring

It has been suggested that residues which are conserved within a subfamily yet not across the complete GPCR family are important for ligand binding [43-45]. To identify ligand binding residues we calculated the sequence conservation (expressed as entropy values) inside- and outside each subfamily for each residue position[38]. A residue score is defined by a combination of both entropy values and reflects the importance of the residue position for ligand binding.

5.2.6. Placement of interaction features

Interaction feature points complementary to the residue properties (acceptor, donor, positive ionizable, negative ionizable and hydrophobic) are positioned in the pocket volume for each rotamer with densities related to the residue score and rotamer likelihood. The density of interaction points is for each rotamer evenly spread over the surface as described by the corresponding interaction geometry [46-50]. All interaction points inside the protein volume or outside the pocket volume are removed resulting in features which are accessible and which can be used to mimic compounds which bind within the 7 TM domain.

5.2.7. Pharmacophore hypothesis generation

Pharmacophore features are generated at regions of high interaction feature density for a given property. It is suggested that residues which are close to each other in space can enhance the importance of those residues in ligand binding [51-54]. By using the interaction feature point density instead of the residue numbers, residues close in space but not necessarily in sequence can enhance each other. A measure for the importance of a feature is subsequently introduced by scoring the ratio between the number of points and the volume of a feature. These are indirectly influenced by rotamer conservation as well as residue score. A consensus hypothesis is introduced by identification of regions of high interaction feature point densities in the overlay of the interaction feature point densities of individual models. A subset of features from each pharmacophore hypothesis is selected to reduce the total number of features and improve virtual screening performance. This subset comprises the top 2 scoring features of each interaction type (acceptor, donor, positive ionizable, negative ionizable and hydrophobic).

5.2.8. Pharmacophore comparison

Pharmacophore hypotheses differ due to the use of different templates on which the models are based and from which the subsequent hypotheses are derived. A measure of robustness is therefore not introduced by comparison of pharmacophore feature locations in different pharmacophore hypotheses, but by analysis of the residues which contribute to all single pharmacophore features. The contribution of a residue to a feature is calculated and depicted in a pie-chart (**Figure 5.2**). The pie-charts of all different pharmacophore features of the different templates can be compared to quantify the robustness and template dependence of the Snooker method to generate pharmacophore hypotheses.

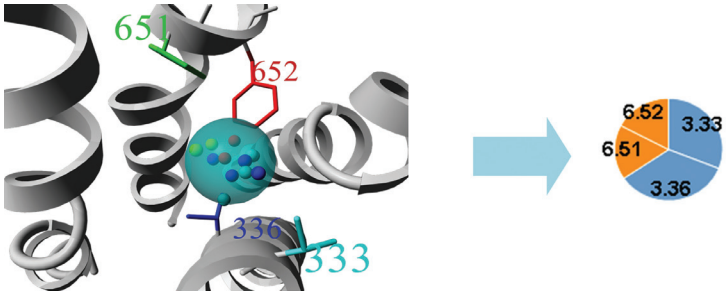


Figure 5.2: Combination of hydrophobic 'interaction points' of four different residues leads to a hydrophobic pharmacophore feature. The contribution of the different residues is visualized in a pie-chart, 4 out of the 12 interaction points can be assigned to residues 3.33 (cyan) and 3.36 (blue) and 2 out of 12 to residues 6.51 (green) and 6.52 (red). Coloring of the pie-charts is according to transmembrane membrane domain from which the residues originate.

5.2.9. Pharmacophore screening

3D conformations of a compound satisfying the pharmacophore constraints are matched to the pharmacophore hypothesis with a minimal root mean square deviation (RMSD) according to the fitting procedure as explained in the Materials and Methods section.

5.2.10. Method validation

Three different experiments were performed to assess the quality of the Snooker pharmacophore hypotheses. First, human beta 2 adrenergic receptor ligands are matched to the Snooker pharmacophore hypothesis and compared to literature supported binding mode hypotheses. Second, an evaluation of whether receptor pharmacophore hypotheses are able to identify the correct ligands with a higher accuracy than 'random' Snooker pharmacophore hypotheses was undertaken. Third, the enrichment of active ligands is tested for multiple class A GPCRs.

5.2.10.1. Retrospective binding mode prediction

To assess whether our approach positions the appropriate features at the correct positions, we have analyzed which residues contribute to pharmacophore features

and evaluated how well the pharmacophore hypotheses perform in a pose prediction experiment. For this purpose we have built pharmacophore hypotheses of the human beta 2 adrenergic receptor and performed a search on a set of known full, partial and inverse agonists as well as antagonists. The RMSD between the poses as reported in literature [55] and the poses as generated after a short energy minimization of the initial pharmacophore matching pose is calculated and reported.

5.2.10.2. Prospective binding mode prediction

Crystal structures of the human DRD3 and human CXCR4 receptor structure co-crystallized with a small molecule have been solved recently [56, 57]. To assess the current status of GPCR modeling, research groups were invited to submit models of those receptors with the bound ligands [58]. We submitted binding mode hypotheses based on Snooker pharmacophores derived from custom made homology models, and optimized with Fleksy [17]. These models were scored on the accuracy of the receptor structure model as well as the predicted binding mode.

5.2.10.3. Target-specific ligand identification

To demonstrate the target specificity of each pharmacophore hypothesis, virtual cross-screens of 15 different GPCR pharmacophore hypotheses and compound sets are performed. The target receptors and families used are listed in **Table 5.1**.

Table 5.1: Targets used in the virtual cross-screen with their corresponding receptor families according to the GPCRDB family classification [59].

Target	Family
5HT7R	Serotonin family
AA2AR	Adenosine family
ADA2B	Alpha adrenoceptors family
ADRB2	Beta adrenoceptors family
AGTR1	Angiotensin family
CLTR1	Cysteinyl leukotriene family
DRD2	Dopamine family
EDNRA	Endothelin family
GASR	Cholecystokinin CCK family
GHSR	Thyrotropin-releasing hormone and secretagogue family
HRH3	Histame family
MCHR1	Melanin-concentrating hormone receptors family
NPY5R	Neuropeptide Y family
OPRM	Opioid family
TA2R	Prostanoid family

The enrichment is calculated for each combination of a compound set and pharmacophore hypothesis. Target specificity can be assumed if the hypothesis and corresponding compound set have better enrichment compared to the average enrichment of all

screens using this same compound set. To show that the final results are not biased by the properties of the active compounds we performed the same experiment for a set of inactive compounds with similar physicochemical properties as the actives.
Library enrichment. Shape constraints are known to improve enrichment in virtual screening [24, 60]. To evaluate the possible performance of the Snooker pharmacophore hypotheses in library enrichment we added a shape constraint to filter the poses as generated in the target specificity experiment and calculated the enrichment of the set of remaining compounds.

5.3. Materials and methods

5.3.1.Template Selection

Three-dimensional coordinates of the template receptors (pdb ids: 1GZM [39], 1L9H [61], 2RH1 [62], 2VT4 [63], 3CAP [64], 3D4S [65], 3DQB [66], 3EML [67]) were obtained from the PDB and structurally aligned using the Needleman and Wunsch algorithm [68] as embedded in YASARA [69] (http:www.yasara.org). Residues which are 5 amino acids apart from the transmembrane helices were removed as were the ligands, waters and other hetero-atoms. Ballesteros and Weinstein numbers [40] of transmembrane helices are: TM1: 1.33-156; TM2: 2.40-2.65; TM3: 3.25-3.51; TM4: 4.43-4.64; TM5: 5.38-5.63; TM6 and 6.37-6.59; TM7: 7.34-7.56.

5.3.2. Homology model

Homology models were built and optimized using WHAT IF [70] and upon completion residues not part of a transmembrane helix were removed.

5.3.3. Rotamer Sampling

The conformational space of all residue side chains is sampled sequentially using the dihedral angles as reported by Lovell et al. [42]. To reduce the number of rotamers we only add a rotamer if it occurs more frequently than 5% in alpha-helices. In order to make our procedure more GPCR-specific we have introduced an additional 50% probability for the rotamer that possesses an average sidechain vector most similar to the average sidechain vector of the available GPCR crystal structures, and we have given a weight of 25% to this particular rotamer in the initial homology model.

5.3.4. Protein- and Pocket Volume Definition

For all residues in our model the Cα-atom position is calculated as well as the average vector between the Cα and the mean coordinate of the sidechain, whilst taking the rotamer distribution into account. This average vector was multiplied by 1.5 and added to the mean Cα position. These calculated coordinates for all residues and 7 dummy points representing the end of the N-terminus and the starts and ends of the extracellular loops are subsequently used in a Delaunay tessellation [71]. The dummy points are positioned

at the end of the vectors starting at the average Cα-atom position of the 4 residues of each helix closest to the extracellular environment, and directed 6Å towards the pocket centre (pocket centre: mean coordinate of the Cα-atoms of residues 3.32, 5.32 and 7.39) and 2Å towards the extracellular side. After removal of all tetrahedra with any edge longer than 8.5Å, the cavities are exposed. All surfaces of which a point 5.0Å along the normal vector is within the original tessellation are marked as potential cavity surfaces. Next, cavity surfaces, which share an edge and have an angle < 0.5π radians between the normal vectors, are merged. Subsequently, the pocket surface is defined as the largest surface of at least 500Å^2 and within 15Å of the pocket centre. Finally, all tetrahedra which have at least one pocket surface and contain no heavy atoms are removed, a procedure which is repeated until all pocket surfaces are part of tetrahedra which contain at least one heavy atom.

5.3.5. Residue Scoring

Receptor families are defined based on sequence homology of the 7 transmembrane helices and entropy scores (reflecting sequence conservation) inside and outside the family are calculated for each position in the multiple sequence alignment [38, 72] according to:

$$E_i = -\sum_{a=1}^{20} F_{ia} \ln(F_{ia}) \quad (1)$$

$$F_{ia} = Number_{ia} / m \quad (2)$$

Where $Number_{ia}$ is the number of sequences with residue type a at alignment position i and m the number of total sequences. Fia is the frequency of residue type a at position i and E_i the entropy of alignment position i. Subsequently one score is calculated for each residue by combining both scores according to equation 3, with E_{in} the entropy inside a cluster and E_{out} the entropy of all sequences not part of the cluster.

$$S_i = \sqrt{\left(E_i^{in}\right)^2 + \left(\ln(20) - E_i^{out}\right)^2} \quad (3)$$

5.3.6. Placing interaction features

Hydrophobic and polar interaction geometries are extracted from Rarey et al. [50]. Charged interactions are described by a cone, the narrow end of which is positioned at the charged center as described by Kumar and Nussinov [73]. The interaction surface is described by the part of the sphere where a vector of 3.5Å, starting at the narrow end of the cone, is within 38° of the vector between the Cα-atom and the corresponding charged centre (**Figure 5.3**).

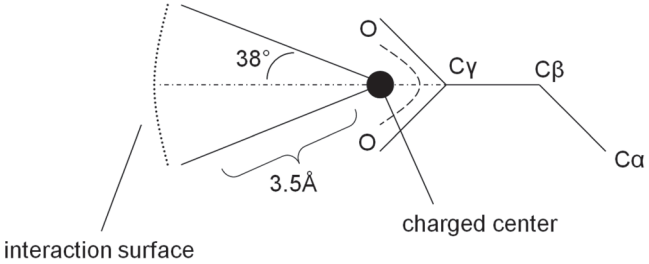


Figure 5.3: Example of a negatively charged interaction geometry.

The final density of interaction points is obtained after sampling a grid on the interaction surfaces. For each rotamer of each residue, sampling starts with the point which is closest to all other interaction points of the same type. Additional points are added iteratively according to:

$$P_{new} = \arg_{x \in P_{left}} \max(D(x, P_{sel}) - D(x, P_{left})) \qquad (4)$$

With Pnew the updated set of interaction points, x a potential point to be added to the selection, Psel the points already selected, Pleft all not yet selected points and D(x,P) the average distance between x and all points in P. Addition of points stops if the number of interaction points has reached:

$$S = N_{basic} \times 1.0/f_i \times R_{ij} \qquad (5)$$

Where f_i is the entropy based residue score of residue i, Rij the rotamer likelihood of rotamer j of residue i and N_{basic} is the initial density for each residue. After the completion of the sampling phase, all points which are not in the pocket volume or in the protein volume are removed.

5.3.7. Pharmacophore hypothesis generation

Pharmacophore features are generated by applying a fuzzy pharmacophore algorithm to all interaction points using an Rc value of 2.5Å [74]. Features are centered at the mean coordinate of all points contributing to the feature and the radius of each feature is set to 2.0Å. The weighted average vector of the centre of all donor and acceptor features towards the residues which contribute to the respective feature is used as a direction vector for the polar feature. The tolerance of this directionality is defined as the minimum of the standard deviation of the vector and ¼π radians. The consensus hypothesis is constructed from the overlay of the interaction feature point densities based on the models using the templates 1GZM, 1L9H, 2RH1, 2VT4, 3CAP, 3D4S, 3DQB and 3EML. Pharmacophore features are ranked according to the number of interaction points per volume. The pharmacophore which is finally used in screening comprises the top 2 ranked features of each interaction type.

5.3.8. Customization of automatically generated pharmacophore hypotheses based on mutagenesis data.

5.3.8.1. EDNRA

Mutagenesis data of the EDNRA receptor suggests a crucial role for Gln3.32 [75]. This residue is not correlated to a pharmacophore feature in the consensus hypothesis due to limited accessibility of this region in models based on some of the templates. However, this region is accessible in the models which have used 3EML and 3CAP as template. Gln3.32 is related to pharmacophore features for the hypotheses based on these templates. Therefore pharmacophore hypotheses based on these two templates are screened for the EDNRA receptor.

5.3.8.2. NPY5R

For the NPY5R the residues Cys3.33 and Cys4.57 were omitted from the Delaunay tesselation. This adjustment is a consequence of the hypothesis that both form a disulfide bridge due to a slight rearrangement of TM3 and TM4 caused by the helix-disturbing residues Pro3.29, Gly4.23 and Pro4.61. Position 5.46 is known to be important for agonism in the aminergic receptors [76] and is also extremely conserved in the NPY5R receptor cluster. Hence, the polar pharmacophore features corresponding to this residue is promoted from rank 3 to rank 2.

5.3.8.3. TA2R

Residue position 7.36 and 7.40 have a remarkably high residue score (data not shown) and are therefore likely to be involved in ligand binding for the tromboxane A2 receptor (TA2R). A modification to the alignment is made for this receptor such that residue 7.36 matches residue 7.35 of the template structure and 7.40 matches 7.39. This effectively results in a kinked helix, as also observed in TM2 of the CXCR4 receptor structure, and positions 7.36 and 7.40 in the pocket. All ligands in the training set contain a carboxyl which seems crucial for the activity. The largest distance to a polar feature is for most ligands within 8-12Å of this carboxyl. We choose to use the pharmacophore based on a beta 2 adrenoceptor template (pdbid: 2RH1) instead of the consensus pharmacophore because the hypothesis derived from the homology model based on the 2RH1 template has a distance of ~12Å between the negative ionizable feature corresponding to Arg7.40 and the acceptor feature near TM5, while this distance is larger in the consensus hypothesis and in the hypotheses based on all other templates.

5.3.8.4. OPRM

Binding mode hypotheses for opioid receptors are described in literature and depict a pharmacophore with a phenolic site deep down a subpocket near TM5, a hydrophobic region near TM5 towards the extracellular loop and an anionic site related to residue 3.32 [77-79]. The pharmacophore hypothesis based on the template 2VT4 seems to mimic

this binding mode best with a positive ionizable and polar feature related to 3.32 and two hydrophobic features near helix V. Therefore we included the pharmacophore hypothesis based on the 2VT4 template as well.

5.3.9. Compound sets

5.3.9.1. Pose prediction

A maximum of 100 3D conformations per molecule was generated for all compounds depicted in Figure 2 of de Graaf et al. [55] using Corina [80] and Cyndi [81].

5.3.9.2. Target specificity

Compounds with activities better than 50nM were retrieved from Chembl02 for the 15 targets and divided into training and test sets. The test set was selected using exclusion sphere clustering on BCI fingerprints and contained the 50 most chemically diverse compounds for each target. The 'fake' active sets of compounds with similar properties to the actives but different architectures were selected based on the same number of positive ionizable and negative ionizable features, acceptors, donors, hydrophobic atoms, aromatic atoms, heavy atoms, rotatable bonds and number of rings. For each known active compound in the test set the most similar (and presumed inactive) compound was chosen from all Chembl02 compounds that have not been tested in GPCR assays. 'Fake' active compounds were selected with emphasis on the same ionizable features and polar groups, and with MACCS fingerprint similarities greater than 0.6 to all known actives. The 10.000 assumed inactive compounds used as decoys were selected from the 10.000 most diverse compounds from all Chembl02 compounds not tested in GPCR assays. This selection was carried out using the diverse molecule component with FCFP_4 fingerprint as implemented in Pipeline Pilot [82]. 3D conformations were generated for all compounds using the procedure described to generate conformations for the pose prediction compound set.

5.3.10. Method validation

5.3.10.1. Retrospective binding mode prediction

The performance of Snooker in reproducing protein-ligand binding hypotheses as reported in literature [55] was investigated for ADRB2. Compound poses matching 5 or more pharmacophore features were minimized using sidechain optimizations and energy minimizations using the Yamber3 forcefield as embedded in YASARA [69]. First, a procedure of optimization of the sidechains of all residues with a distance less than 10Å from the ligand pose using SCWALL [83] as implemented in YASARA followed by an energy minimization with fixed carbon atoms of the helix endings is performed twice. After this all sidechains with a distance less than 8Å from the ligand are again optimized using SCWALL. Finally, the complex is energy minimized without constraints using the Yamber3 forcefield. The resulting poses are then compared to corresponding reference

poses reported by de Graaf et al. [55], by calculating the RMS between all atoms within 13Å of the Cα-atom of Ser5.46 in the reference structures and their counterparts in the pharmacophore guided poses. Although the reference structures are obtained via a docking experiment in customized receptor models, the validity of these models is supported by their high similarity to the recently elucidated structures of the homologues ADRB1 receptor with several ligands [84].

5.3.10.2. Prospective binding mode prediction

PDB entry 2RH1 was used as a template to build the homology models. The structure was cleaned, the lysozyme protein was removed and the bound ligands retained. Residues in the loop between TM5 and TM6 (residues 218-317) were discarded in the modeling process for the DRD3 receptor. For CXCR4, residues in the loop between TM6 and TM7 (residues 267-276) were discarded and the loop between TM4-TM5 was deleted (residue Y174-A198) and replaced with the loop TM4-TM5 from PDB entry 3EML (AA2AR_HUMAN, residues N144-N175). Homology modeling was performed using the Yasara program and its built-in modeling algorithm [85]. Subsequently, the initial models were manually refined and subjected to a final energy minimization step. Snooker pharmacophores were generated based on the custom built homology models by the procedure described in this manuscript. Finally, compound matches in all 5 feature pharmacophores for eticlopride in the DRD3 receptor and in all 4 feature pharmacophore for the CXCR4 receptor were optimized and scored using Fleksy [17].

5.3.10.3. Target specificity

The performance of Snooker with regard to target specificity was tested in a cross-screening exercise. A virtual cross-screen of 15 pharmacophore hypotheses of different GPCRs and 10000 assumed inactive compounds plus 750 active compounds (50 for each GPCR) was therefore performed. After the virtual screen, compounds were first ranked according to the number of fitted features, and subsequently by their fit value. For each set of 50 actives and 10000 assumed inactives, the area under the semi-logarithmic receiver operating curve (pROC AUC) was calculated [86]. Finally, the 15 different pharmacophore hypotheses were ranked for each compound set using the pROC AUC. Since the pROC AUC emphasizes early enrichment we have chosen this value instead of the normal ROC AUC. To show that results are not biased towards the properties of the active compounds, 15 fake sets of active compounds were created and screened together with the 10000 assumed inactives. To construct the fake sets, one compound which has not been tested in a GPCR assay and possesses similar properties to an active was selected per active molecule. The pROC AUC and ranking is subsequently calculated for each pharmacophore hypothesis and compound as described before.

5.3.10.4. Library enrichment

From all training set compounds poses we removed those which have at least one atom within 2.0Å of the backbone of the homology model. The pose which had the best volume overlap (calculated as the smallest average tanimoto shape distance) with all filtered training set poses was defined as the reference. Next, the volume overlap between the reference and all poses of the 10050 compounds (10000 compounds not tested in a GPCR assay and 50 actives for the GPCR of interest) which match the pharmacophore were calculated. A value for the volume overlap (as tanimoto shape distance) was set at the maximal enrichment of training set actives versus decoys and including at least 25% of training set actives. Subsequently, a shape cutoff was defined by adding 0.03 to this value. All poses of the test and decoy set with a tanimoto shape distance larger than the cutoff were removed and the remaining compounds were ranked according to the number of matching pharmacophore features, the number of actives in the training set which hit this same pharmacophore and finally the fit value. Lastly, the AUC, pROC AUC [86], enrichment at 0.5%, 1%, 2% and 5% of the compound set were calculated [87].

5.3.11. Pharmacophore screening.

Pharmacophore searches were performed using custom code implemented in python and making use of RDKit [88]. First, ligand atoms were typed using the built-in definitions of RDKit. Second, all matches between pharmacophore features and ligand features were listed after which all possible matches between a ligand and a pharmacophore were calculated. Third, all ligand atoms which match features were transformed to minimize the RMS between the ligand atoms and the centre of the corresponding pharmacophore feature. Fourth, the distance from each atom to the centre of the feature it should match was calculated, and if this was larger than the pharmacophore radius, the weight of this pair was increased in the next transformation. This procedure is repeated three times and a hit was defined when the average RMS was smaller than the average pharmacophore radius, and when all ligand atoms were within 105% of the radius of the pharmacophore feature.

5.4. Results and discussion.

5.4.1. Retrospective binding mode prediction

The contributions of the different residues to the pharmacophore hypotheses deduced from the ADRB2 models derived from 8 different template structures as well as a consensus are depicted in **Figure 5.3**. The consensus pharmacophore hypothesis is derived from the overlay of the models based on the 8 different templates and contains the most robust features related to the residues identified as being important for ligand binding by multiple sequence alignment (MSA) analysis. Represented are for example known important interactions like the positive ionizable interaction of Asp3.32, the hydrophobic contact with Val3.33, the polar interactions with Asn7.39 & Asp3.32 and with Ser5.42 & Ser5.46 [76]. The overlay of the 8 models represents all possible interactions hypothesized by the individual models. A weighted average of all these possible interactions defines which single interactions are considered most important based on MSA analysis and robust as indicated by their presence in the majority of the models. The consensus hypothesis is newly created by revisiting the interaction feature points and recalculation of the densities, thus pharmacophore features. For example, Ser5.46 has been shown to be important for the binding of agonists [76], but has only limited accessibility in the models based on the aminergic templates, probably due to the antagonist conformation in which the aminergic receptors have been crystalized. This same area of the receptor is much more accessible in the models based on the opsin and rhodopsin templates and promotes the use of models based on these templates to construct the consensus pharmacophore hypothesis. Ser5.46 is very well represented in the consensus pharmacophore hypothesis due to the inclusion of the rhodopsin templates (**Figure 5.4**).

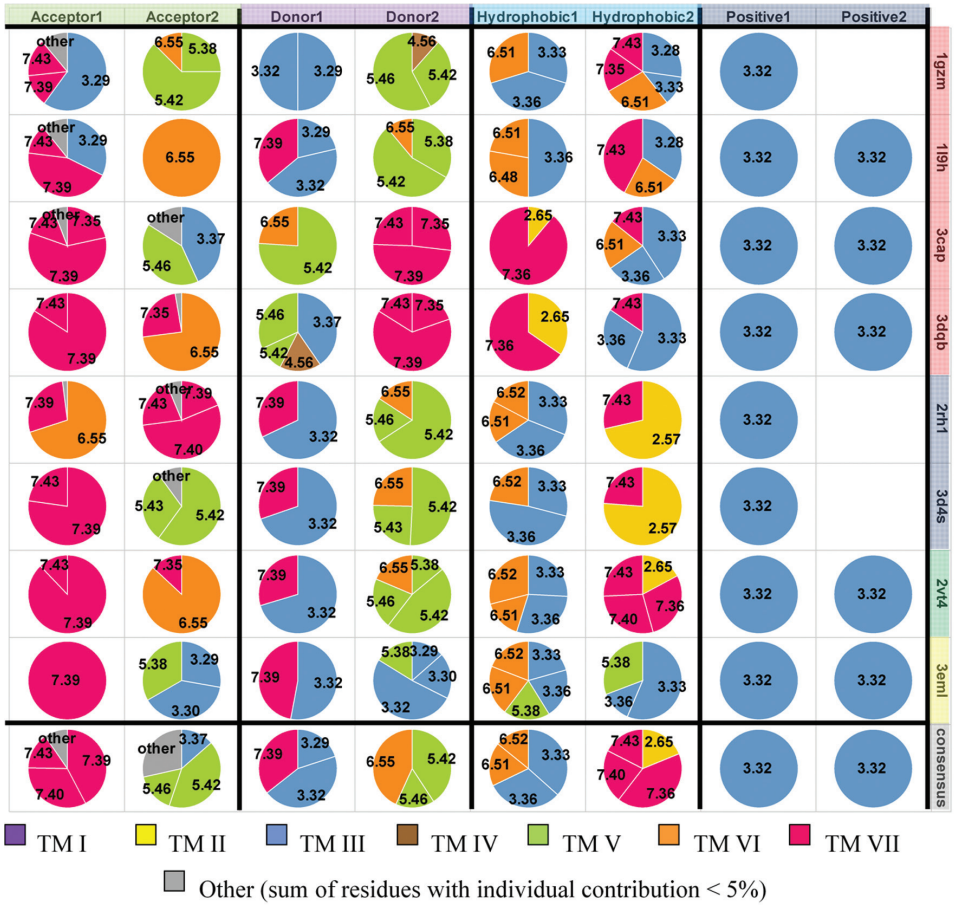


Figure 5.4: Contribution of the different residues to the different pharmacophore features (horizontal) for the pharmacophore hypotheses derived from the ADRB2 models based on the different templates as well as the consensus pharmacophore hypothesis (vertical). Coloring of the pie-charts is according to the transmembrane domain in which the residue is positioned. Templates corresponding to crystal structures of beta aminergic, bovine rhodopsin and the adenosine A2A receptors are colored blue, red and green, respectively.

The pharmacophore guided poses of 16 compounds matching a consensus (sub) pharmacophore hypothesis of 5 or more features are shown in **Figure 5.5** together with the reference as defined by de Graaf et al. [55]. The depicted poses are the most similar poses to the reference amongst all possible poses which satisfy a pharmacophore of at least 5 features. Since evolutionary pressure is mainly driven by endogenous agonists [89] and the fact that the Snooker procedure assigns weights to residues based on calculated entropies (reflecting evolutionary pressure) from a multiple sequence alignment comprising multiple species, it is to be expected that the method is more biased towards agonism than towards antagonism. This and the fewer specific contacts for antagonists

as suggested by mutagenesis experiments [90] might explain the relatively high number of antagonist experiments (9 out of 12 antagonists [red] compared to 1 out of 14 agonists [green]) for which no pose could be produced. 13 of the 16 ligands have at least one pose comparable to the reference (heavy atom root mean square deviation (RMSD) <= 2.0Å). This indicates that Snooker pharmacophores represent the correct protein-ligand interactions for the majority of ADRB2 ligands which do fit a (sub)pharmacophore of at least 5 features.

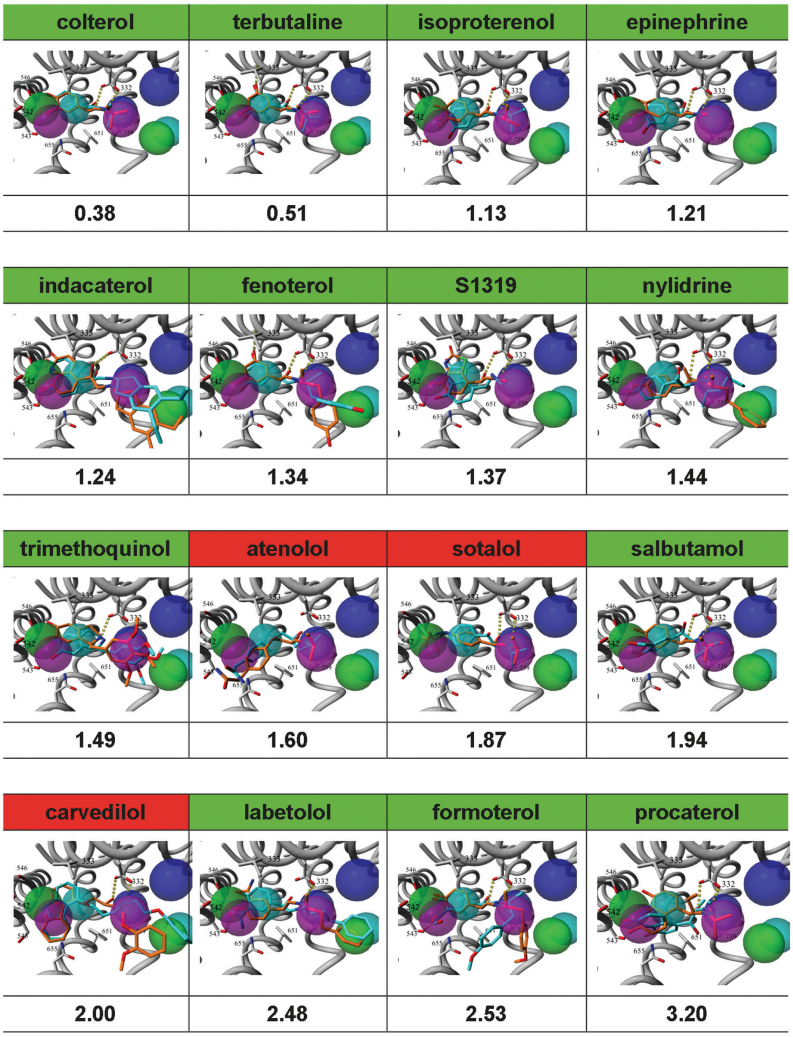


Figure 5.5: Receptor structure of the reference structure (gray), reference pose (orange), most similar predicted pose (cyan) and structure-based consensus pharmacophore hypothesis of the human beta 2 adrenergic receptor. Pharmacophore features are colored cyan (hydrophobic), green (acceptor), magenta (donor) and blue (positive ionizable). Ligand names are colored red for antagonist/inverse agonist and green for agonist. The heavy atom root mean square deviation (RMSD) of each pose is indicated below each figure.

5.4.2. Prospective binding mode prediction

Binding mode hypotheses largely based on Snooker pharmacophores are generated for the human DRD3 and CXCR4 receptor and submitted for evaluation in the GPCR dock 2010 assessment [58]. Utilizing custom made homology models, Snooker pharmacophores and low resolution binding modes were generated and optimized by Fleksy. The dopamine DRD3 compound, eticlopride, fitted dominantly into one five feature pharmacophore and was consistently predicted in a binding mode similar to the pose depicted in **Figure 5.5**. Based on the five feature pharmacophore used to generate these poses, interactions with Asp3.32, His6.55, Tyr7.35, Thr7.39, Tyr7.43, Val3.33, Val5.39 and Phe6.51 can be assumed. The final poses as optimized by Fleksy indeed represent those interactions and all 5 submitted models ranked in the top 10 in the DRD3 assessment. The best model correctly predicted 36/65 atomic contacts and 12/15 residue contacts [58].

The default CXCR4 model did not show a pocket volume for the positioning of pharmacophore features due to bulky and inward directed residues in the upper part of the transmembrane domain. To generate a pocket volume the minimal edge length required for the removal of tetrahedra was reduced from 8.5Å to 8.0Å and 7.5Å. This resulted in two different subpockets in the minor binding pocket between TM2, TM3, and TM7[57]. Pharmacophores related to both subpockets were generated and screened resulting in two distinct sets of possible poses. IT1t poses in the first subpocket 1 correlated to interactions with residues Glu7.39, Thr3.33, Phe2.57, Leu 2.61, Tyr3.32, Leu3.36 & Phe7.43 and poses in the second subpocket 2 represented interactions of the small molecule IT1t to residues Glu7.39, Tyr3.32, His3.29, Val3.28 & Leu2.61. Since the pharmacophore features from the first subpocket 1 are based on a larger number of residues originating from the transmembrane domain we optimized and submitted compound poses in this pocket for evaluation and ranked 14th, 17th, 28th, 29th and 33th out of 103 predictions in the GPCR dock 2010 assessment [58].

The pose prediction experiment of ADRB2 already indicated that antagonists and inverse agonists are likely to have less specific contacts as agonists. This and the observation that the co-crystalized peptide ligand in the CXCR4 crystal structure does occupy a large part of the major pocket indicates that other small molecules and agonists in particular might bind somewhere other than IT1t. Interestingly, retrospective analysis of the binding modes obtained in minor subpocket 2 (which was not submitted for the GPCR DOCK 2010 assessment) show structures which resemble the CXCR4-IT1t crystal structure with interactions to residues Glu7.39, Tyr3.32, His3.29, Val3.28 & Leu2.61 (**Figure 5.6B,D**). Important interactions with W2.60 and D2.63 are however not predicted in these models, as the T2.56XP2.59 induced kink in TM2 was not correctly predicted based on the available set of GPCR crystal structure templates [58].

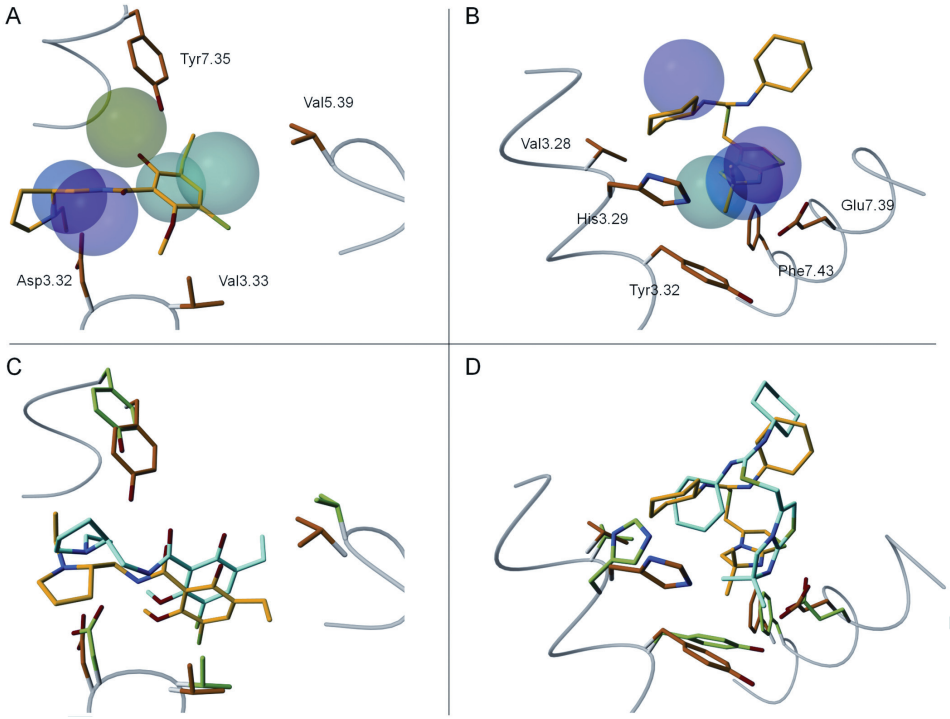


Figure 5.6: Binding mode predictions of eticlopride in the human DRD3 receptor and 1t in the CXCR4 receptor. **A**: Human DRD3 pharmacophore and matching eticlopride pose. **B:** Human CXCR4 pharmacophore and matching IT1t pose. **C:** Optimized eticlopride pose and corresponding receptor structure in yellow and orange and crystal structure pose and receptor structure in green and cyan. **D:** Optimized IT1t pose and corresponding receptor structure in yellow and orange and crystal structure pose and receptor structure in green and cyan.

5.4.3. Identification of receptor-specific ligands and ligand binding residues in cross-screen.

Consensus pharmacophore hypotheses are generated for 15 targets using default Snooker settings, as described in the Methods section. The residue contributions to the pharmacophore hypotheses for the different models are depicted in **Figure 5.7**. The significance of each feature for the identification of known actives for a model is indicated below the pie-chart. This significance has been calculated as the fraction of pharmacophores which contain the particular feature and match a conformation of a known active. Site-directed mutagenesis studies confirm that most of the pharmacophore features in the Snooker models relate to ligand binding residues. Most positively ionizable (and/or H-bond donor) pharmacophore features are associated with negatively charged residues which are determined to be essential for ligand binding to bioaminergic receptors ADA2B, ADRB2, DRD2, HRH3, and 5HT7R (D3.32) [76], EDNRA (D7.35) [75], GHSR (E3.33) [91], MCHR1 (D3.32) [92] and OPRM (D3.32) [78]. Likewise, negatively charged (and/or

acceptor) features correspond to positively charged residues shown to be involved in ligand binding in AGTR1 (K5.42) [93], EDNRA (R6.55) [75], GHSR (R6.55) [91] and TA2R (R7.40) [94]. In addition, polar residues at positions T3.36/N6.55/S7.42 in AA2AR [67, 95], 5.42/5.43/5.46 in the bioaminergic receptors [76] and Q3.22 in EDNRA [75] are associated with donor and acceptor features and indeed play key-roles in ligand binding for these receptors. The Snooker approach clearly allows an automated and fully protein-based construction of experimentally supported pharmacophore models. Interestingly, valid pharmacophore models could not only be built for receptors with existing crystal structures (ADRB2 and AA2AR) and receptors which are related to crystallized GPCRs (5HT7R, ADA2B, DRD2, HRH3), but also for receptors with low sequence similarity to GPCR crystal structure templates (AGTR1, CLTR1, EDNRA, GASR, NPY5R, GHSR, MCHR1, OPRM, TA2R).

Despite the fact that all template structures have ligands with key interactions to helix V, all features of the GHSR receptor are almost exclusively related to residues in helix III, VI and VII. Site directed mutagenesis experiments confirm the importance of several residues (Q3.33/F6.51/R6.55/P6.58/N7.35) of helix III, VI and VII and the relative unimportance of helix V (M5.39/V5.42/S5.43/F5.46) for ligand binding [91], illustrating the ability of Snooker to presume different binding pockets by using similar template structures. The features of the CLTR1 pharmacophore hypothesis relate only to residues originating from helices III, VI and VII. This suggests a different binding mode for ligands of this receptor compared to e.g. the amine receptors (5HT7R, ADA2B, ADRB2, DRD2 and HRH3) which do have polar interactions with helix V [76]. The values below the features indicate the importance of those features in retrieving active compounds and it is remarkable to see that positive ionizable features play a crucial role in retrieving active compounds when present (as witnessed by the relatively high values). However, due to the limited number of negative ionizable residues and therefore positive ionizable features in the pocket, are often two positive ionizable features related to the same residue (AA2AR, ADA2B, ADRB2, AGTR1, HRH3, OPRM). Subpharmacophores which are identical except for the positive ionizable features are therefore likely to retrieve similar compounds by assuming almost identical ligand poses. The absence of a feature in pharmacophores retrieving active compounds gives the opportunity to drop such a feature in a pharmacophore for virtual library screening and will most likely increase the hit rate.

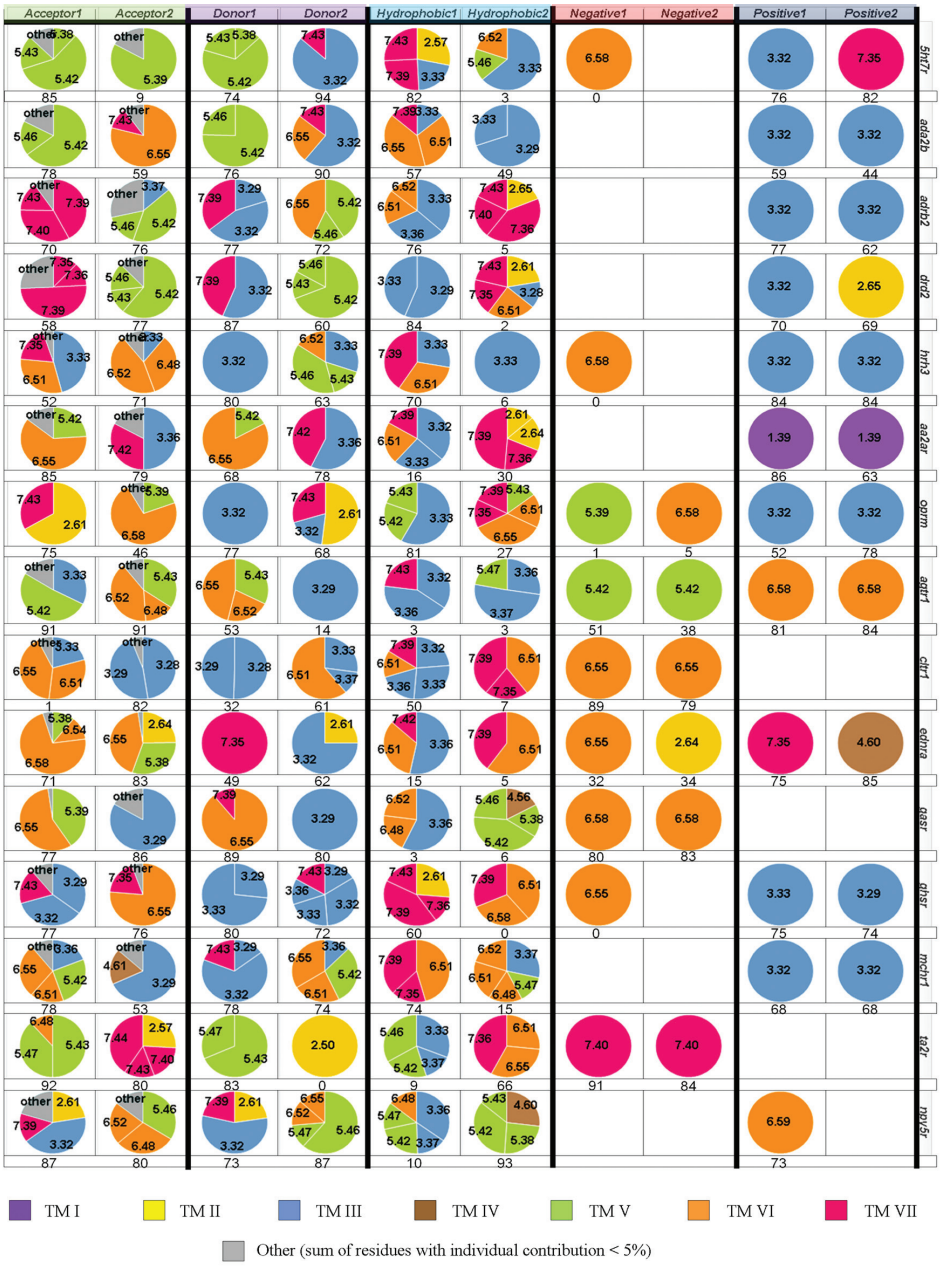


Figure 5.7: Pie-charts visualizing the contributions of the different residues to pharmacophore hypotheses for 15 different GPCRs. The percentage of recognized active compounds by a subpharmacophore which contain a feature is indicated below the pie-chart of that particular feature. In a scenario where 60 out of the 100 actives are retrieved, and 30 of the actives are matched in a pharmacophore where feature X is involved, the number for feature X will be 50.

To validate the target-specificity of the generated pharmacophore hypotheses we performed a cross-screening experiment on 15 sets of active and decoy compounds using the 15 corresponding pharmacophore hypotheses. The enrichment of all sets of active compounds is calculated for all pharmacophore hypotheses. Enrichment values of the different hypotheses are ranked for each set of active compounds. **Figure 5.8** shows the rank of the pharmacophore hypothesis of the target receptor which corresponds to the compound set of actives. The target consensus pharmacophore hypothesis ranks as one of the best 3 hypotheses for 10 of the 15 receptors.

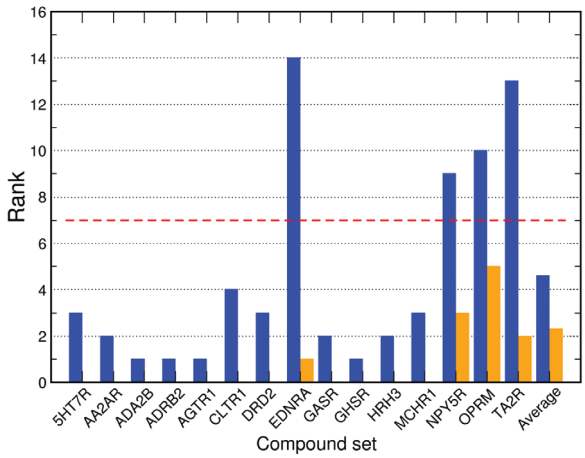


Figure 5.8: The rank in a list of 15 pharmacophore hypotheses at which a target pharmacophore hypothesis enriches a compound set of active molecules for that target. Blue indicates the rank of the consensus models and orange the rank of the adjusted models (see Materials and Methods). The average rank of the screen of the "fake" sets of actives is indicated with the red dashed line.

A shortcoming of the automatic homology modeling procedure is the accurate prediction of helix structure, for example when kinks are present, as observed in TM2 of the recently elucidated CXCR4 receptor structure [57]. Performance of the consensus pharmacophore hypotheses in the cross screening experiment is poor for 4 compound sets. Mutagenesis data and multiple sequence analysis [38, 72] indicate that minor model adjustments are required for those receptors. Gln3.32 is known to be important for ligand binding in the EDNRA receptor [75] and is not related to a pharmacophore feature in the original pharmacophore hypothesis. The hypotheses based on the 3CAP and 3EML template is tested for the EDNRA receptor because Gln3.32 correlates to a pharmacophore feature in these hypotheses. The modifications made to the NPY5R, TA2R and OPRM receptor are, respectively, the adjustment of the template due to an assumed sulfide bridge and the inclusion of the third ranked polar feature instead of the second ranked polar feature, adjustment of the multiple sequence alignment and selection of a template to enable exposure of residues which are likely to be ligand binding, and selection of a template due to correspondence of the resulting pharmacophore hypothesis with literature [77-

79] (see Methods for details). The original consensus pharmacophore of those receptors show a reasonable amount of similarity to the modified pharmacophores and show major differences by 2-3 features (data not shown).
Three of the manually adjusted pharmacophore hypotheses retrieve their corresponding compound set amongst the best 3 hypotheses in the earlier mentioned cross-screen. Thus, as in many other modeling procedures, knowledge-based input can also improve Snooker pharmacophore hypotheses [52, 55], but such input is not a prerequisite for this method.

All 15 generated pharmacophore hypotheses possess at least 2 acceptors, 2 donors and 2 hydrophobic features and some are complemented with a maximum of 2 positive and/or negative ionizable features. The presence of ionizable features in hypotheses could potentially lead to biased enrichment – the requirement for corresponding features on ligands might function as an effective 2D filter, independent of positions in the hypotheses. To address this problem, fake active sets of compounds with chemical properties similar to the active molecules were cross-screened to determine the target specificity accomplished by the correct spatial arrangement of features. The average rank of 7.0 for the pharmacophore hypotheses corresponding to the fake active sets (compared to 2.3 for the true actives) further indicates that features have indeed the correct spatial arrangement. Ideally, compound sets should be best recognized by the pharmacophore hypothesis based on the homology model of the receptor at which the compounds are active. However, polypharmacology against GPCRs is widely accepted and even occurs in known antipsychotics and antidepressants. The selected compound sets include the 50 structurally most diverse compounds for each target and possibly contain a mix of partial and full agonists, antagonists and inverse agonists. Hierarchical clustering on compound fingerprints shows the similarity of compounds selected for the different targets. Some of the compounds show activity across multiple target receptors (Chembl02 compound id. 2214 has reported activity on the 5HT7R and DRD2 receptor, compound id. 1989 has reported activity on the ADA2B and DRD2 receptor and compound id. 27397 has reported activity on the ADA2B and ADRB2 receptor). Although active compounds are only included in the test set of only one receptor, they are successfully retrieved by the pharmacophore hypotheses of the other targets on which they have reported activity (data not shown). Cross-activity inhibits the selection of sets of compounds uniquely active at only one receptor and it is of course possible that cross activity exists, especially within receptor sub-families, which is not yet reported in the literature. Snooker allows for the virtual screening of compounds against pharmacophores of all class A GPCRs. The outcome of such a virtual screen can be used to predict bioactivity profiles, design multi-target drugs, or to select a panel of GPCRs on which a compound should be tested to prohibit possible side effects at a later stage.

5.4.4. Library enrichment

The ultimate goal of 3D pharmacophore hypotheses is to predict the conformation of all and only active ligands exactly as observed in the crystal structure. Shape constraints are often added to pharmacophore hypotheses to improve the enrichment of active molecules in virtual screenings of compound libraries [24, 60]. Since only limited enrichment is observed in the cross-screening experiments we introduce a shape constraint derived from a reference pose of an active compound in the enrichment experiment to optimize enrichment factors. The reference pose is automatically extracted from a training set of active compounds and used to filter the decoys and test set on shape similarity to this reference pose (see Materials and Methods). The number of actives, decoys and enrichment factors for the filtered and unfiltered sets are listed in **Table 5.2**.

Table 5.2: Retrospective virtual screening accuracies of 15 compound sets using pharmacophores with and without shape constraints.

Receptor	Actives (unfiltered)	decoys (unfiltered)	shape cutoff[a]	actives (filtered)	decoys (filtered)	EF (unfiltered)[b]	EF (filtered)[c]
5HT7R	16	3023	0.560	5	567	1.06	1.76
AA2AR	25	4170	0.555	9	884	1.20	2.04
ADA2B	30	5232	0.434	12	38	1.15	63.16
ADRB2	38	2253	0.576	21	452	3.37	9.29
AGTR1	22	1961	0.635	9	350	2.24	5.14
CLTR1	13	2823	0.514	2	20	0.92	20.00
DRD2	30	2918	0.560	12	845	2.06	2.84
EDNRA	44	5721	0.578	24	1917	1.54	2.50
GASR	11	1444	0.844	11	1230	1.52	1.79
GHSR	48	4439	0.588	17	508	2.16	6.69
HRH3	37	4762	0.570	19	1061	1.55	3.58
MCHR1	36	4386	0.603	13	802	1.64	3.24
NPY5R	14	1925	0.618	4	762	1.45	1.05
OPRM	24	3610	0.668	5	678	1.33	1.47
TA2R	25	675	0.538	5	1	7.41	1000.0

[a] Shape cutoff used to filter the test and decoy set (**paragraph 5.3.10.4**).
[b] Enrichment factor (EF): ratio of "filtered actives / filtered decoys" to "all actives / all decoys".
[c] Enrichment factor (EF): ratio of "unfiltered actives / unfiltered decoys" to "all actives / all decoys".

On average the shape constraint improves enrichment values by a factor ~2.0. The enrichment values for those targets with the most stringent shape criteria (ADA2B, CLTR1, TA2R) are improved best. In the case of ADA2B this is due to the small size of the reference molecule as well as ADA2B active compounds. Larger compounds have by definition a higher chance to fit a pharmacophore without shape constraints and are removed by the addition of this shape constraint. As a result, the average molecular weight of the filtered sets is reduced from 201 to 165 and 279 to 173 for the actives and decoys respectively. For TA2R the shape constraint selects almost exclusively poses which fit the pharmacophore consisting of features 1,2,6,7 and 8. This pharmacophore is rarely

present in poses generated for decoy compounds. A result of the stringent cutoff is that all remaining compounds after filtering contain the same scaffold. However it should be noted that an acceptor feature and a correct orientation of the varying R-group is required to fulfill the pharmacophore criteria and shape constraint. For CLTR1 very few actives or decoys were selected. Retrospective analysis of the training set shows a high similarity among the compounds in this training set explaining the stringent cutoff value and low number of selected compounds. Both selected CLTR1 actives also show relatively high similarity to the compounds in the training set, indicating that the automatic selection of a reference and cutoff value can sometimes result in high enrichment values but with less novelty in the resulting compound sets due to a suboptimal training set. Most shape filtered sets do however show chemically diverse active compounds as desired in typical virtual screening experiments. The training set of active compounds is used to derive the reference shape but also to rank the compounds which match a pharmacophore. Pharmacophores are therefore first ranked on the number of features and next on the number of training compounds which match the pharmacophore. Compounds are subsequently ordered on the pharmacophore in which they match and the fitvalue which is obtained in this pharmacophore. Using this approach we calculated 6 different performance measures for the enrichment and these are listed in **Table 5.3**. The high early enrichment values (at EF 0.5% and 1.0%) as compared to the enrichment values in **Table 5.2** indicate that this ranking improves the result of the virtual screening.

Table 5.3: Retrospective virtual screening accuracies of 15 compound sets using Snooker pharmacophores and shape constraints. Enrichment factors (EF) above 3 and 10 are colored orange and green, respectively.

Receptor	AUC[a]	pROC AUC[b]	EF[c] at 0.5%	EF[c] at 1.0%	EF[c] at 2.0%	EF[c] at 5.0%
5HT7R	0.53	0.57	12	6	4	2
AA2AR	0.55	0.53	0	2	1	2
ADA2B	0.61	0.98	44	22	11	5
ADRB2	0.69	1.04	16	14	14	8
AGTR1	0.58	0.71	16	12	9	4
CLTR1	0.52	0.56	8	4	2	2
DRD2	0.59	0.76	16	10	8	5
EDNRA	0.66	0.90	16	8	5	5
GASR	0.56	0.62	4	10	5	3
GHSR	0.65	0.96	16	14	10	7
HRH3	0.64	0.82	8	6	4	6
MCHR1	0.60	0.72	8	4	6	4
NPY5R	0.50	0.43	0	0	1	1
OPRM	0.52	0.50	0	2	3	2
TA2R	0.54	0.71	16	8	4	2

[a]Area under the ROC curve;
[b]Area under the semi-logarithmic curve;
[c]Enrichment factor (EF): the ratio of true positive rates to false positive rates at increasing false positive rates (0.5%, 1%, 2% and 5%).

Virtual screening methods should enrich active ligands at least 10 fold to obtain a reasonable chance of finding a true hit [96]. According to this criterion, Snooker shows early enrichment for 8 and 6 of the 15 receptors at 0.5% and 1% false positive rates. Interestingly, high virtual screening enrichments are obtained not only for bioaminergic receptors (ADRB2, 5HT7R, ADA2B, DRD2), but also pharmacophore models based on receptors with low sequence similarity to GPCR crystal structures (AGTR1, EDNRA, GASR, GHSR, TA2R) yield high early enrichment results. This shows that Snooker is not necessarily dependent on the availability of high resolution structural data, and demonstrates the strength of the pharmacophore modeling approach. Although enrichment is mainly achieved by the additional shape restraints, this method results in a binding model of the compounds which is likely to reflect the true interaction of the compound with the receptor. Such information can be very useful in the design of experiments and compound optimization after the discovery of a new active compound. The relatively poor performance of the AA2AR pharmacophore hypothesis, partly based on structural information of the AA2AR crystal structure can be explained by the fact that the ligand binding site is for a large part located between the extracellular loops, currently not included in Snooker pharmacophore models. It should furthermore be noted that the test sets of 50 active compounds consist of compounds with different scaffolds and possibly different binding modes [97], explaining the generally low global virtual screening results (AUC). Such a diverse set of active molecules is possibly more difficult to describe with a combination of a single pharmacophore hypothesis and shape restraint and results therefore usually in less satisfying enrichments. The definition of multiple combinations of pharmacophores and shape restraints based e.g. overlays of different compound series available in the training set might improve enrichment scores but would likely limit the structural diversity of hits.

5.5. Conclusion

In this paper, we present Snooker, a new structure-based approach to generate low resolution pharmacophore hypotheses for class A GPCRs. We show that Snooker generates ADRB2 pharmacophore hypotheses which retrieve 3/12 antagonists-inverse agonists and 13/14 agonists and assumes the correct binding mode for 3 and 10 compounds, respectively. All 5 submitted eticlopride binding mode predictions in the human DRD3 receptor, which are largely based on the eticlopride matches in Snooker pharmacophore, showed to be in the top 10 of all submitted models in the GPCR dock 2010 assessment. The automated and fully protein-based construction of pharmacophore hypotheses is in line with experimental site-directed mutagenesis data on essential ligand binding residues for a diverse set of 15 class A GPCRs. For several of the more difficult targets, the default Snooker settings can be adjusted with target-based knowledge to obtain good results. Interestingly, valid pharmacophore models were built for not only 2 receptors with known crystal structures and 4 related receptors but also for 9 receptors with low sequence similarity to GPCR crystal structure templates. A virtual screening experiment using the Snooker pharmacophore hypotheses in combination with a shape constraint resulted in >10 fold enriched compound sets for 8 out of 15 targets. As such, the method is suitable to design focused compound libraries targeting a small subfamily of class A GPCRs.

References

1. Klabunde, T. and G. Hessler, *Drug design strategies for targeting G-protein-coupled receptors*. ChemBioChem, 2002. **3**(10): p. 928-44.
2. Hopkins, A.L. and C.R. Groom, *The druggable genome*. Nat Rev Drug Discov, 2002. **1**(9): p. 727-30.
3. Franke, L., et al., *Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors*. J Med Chem, 2005. **48**(22): p. 6997-7004.
4. Marrero-Ponce, Y., et al., *Ligand-based virtual screening and in silico design of new antimalarial compounds using nonstochastic and stochastic total and atom-type quadratic maps*. J Chem Inf Model, 2005. **45**(4): p. 1082-100.
5. Schuster, D., et al., *The discovery of new 11beta-hydroxysteroid dehydrogenase type 1 inhibitors by common feature pharmacophore modeling and virtual screening.* J Med Chem, 2006. **49**(12): p. 3454-66.
6. Evers, A., et al., *Virtual screening of biogenic amine-binding G-protein coupled receptors: comparative evaluation of protein- and ligand-based virtual screening protocols.* J Med Chem, 2005. **48**(17): p. 5448-65.
7. Schneider, G. and M. Nettekoven, *Ligand-based combinatorial design of selective purinergic receptor (A2A) antagonists using self-organizing maps*. J Comb Chem, 2003. **5**(3): p. 233-7.
8. Oloff, S., R.B. Mailman, and A. Tropsha, *Application of validated QSAR models of D1 dopaminergic antagonists for database mining.* J Med Chem, 2005. **48**(23): p. 7322-32.
9. Zhang, Q. and I. Muegge, *Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring*. J Med Chem, 2006. **49**(5): p. 1536-48.
10. Oprea, T.I. and H. Matter, *Integrating virtual screening in lead discovery.* Curr Opin Chem Biol, 2004. **8**(4): p. 349-58.
11. Congreve, M., C.W. Murray, and T.L. Blundell, S*tructural biology and drug discovery.* Drug Discov Today, 2005. **10**(13): p. 895-907.
12. Jorgensen, W.L., *The many roles of computation in drug discovery.* Science, 2004. **303**(5665): p. 1813-8.
13. Hou, T. and X. Xu, *Recent development and application of virtual screening in drug discovery: an overview*. Curr Pharm Des, 2004. **10**(9): p. 1011-33.
14. de Graaf, C. and D. Rognan, *Customizing G Protein-coupled receptor models for structure-based virtual screening*. Curr Pharm Des, 2009. **15**(35): p. 4026-48.
15. Kolb, P., et al., *Structure-based discovery of beta2-adrenergic receptor ligands.* Proc Natl Acad Sci U S A, 2009. **106**(16): p. 6843-8.
16. Katritch, V., et al., *Structure-based discovery of novel chemotypes for adenosine A(2A) receptor antagonists.* J Med Chem, 2010. **53**(4): p. 1799-809.
17. Nabuurs, S.B., M. Wagener, and J. de Vlieg, *A flexible approach to induced fit docking.* J Med Chem, 2007. **50**(26): p. 6507-18.
18. Crossley, R., T*he design of screening libraries targeted at G-protein coupled receptors.* Curr Top Med Chem, 2004. **4**(6): p. 581-8.
19. Salo, O.M., et al., *Virtual screening of novel CB2 ligands using a comparative model of the human cannabinoid CB2 receptor.* J Med Chem, 2005. **48**(23): p. 7166-71.
20. Kenyon, V., et al., *Novel human lipoxygenase inhibitors discovered using virtual screening with homology models.* J Med Chem, 2006. **49**(4): p. 1356-63.
21. Bissantz, C., et al., *Protein-based virtual screening of chemical databases. II. Are homology models of G-Protein Coupled Receptors suitable targets?* Proteins, 2003. **50**(1): p. 5-25.
22. Bissantz, C., A. Logean, and D. Rognan, *High-throughput modeling of human G-protein coupled receptors: amino acid sequence alignment, three-dimensional model building, and receptor library screening.* J Chem Inf Comput Sci, 2004. **44**(3): p. 1162-76.
23. Kratochwil, N.A., et al., *An automated system for the analysis of G protein-coupled receptor transmembrane binding pockets: alignment, receptor-based pharmacophores, and their application*. J Chem Inf Model, 2005. **45**(5): p. 1324-36.
24. Klabunde, T., C. Giegerich, and A. Evers, *Sequence-derived three-dimensional pharmacophore models for G-protein-coupled receptors and their application in virtual screening*. J Med Chem, 2009. **52**(9): p. 2923-32.

25. Tsai, K.C., et al., *Discovery of a novel family of SARS-CoV protease inhibitors by virtual screening and 3D-QSAR studies.* J Med Chem, 2006. **49**(12): p. 3485-95.
26. Bologa, C.G., et al., *Virtual and biomolecular screening converge on a selective agonist for GPR30.* Nat Chem Biol, 2006. **2**(4): p. 207-12.
27. Overington, J.P., B. Al-Lazikani, and A.L. Hopkins, *How many drug targets are there?* Nat Rev Drug Discov, 2006. **5**(12): p. 993-6.
28. Knox, C., et al., *DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.* Nucleic Acids Res, 2011. **39**(Database issue): p. D1035-41.
29. Chen, X., Y. Lin, and M.K. Gilson, *The binding database: overview and user's guide.* Biopolymers, 2001. **61**(2): p. 127-41.
30. Chen, X., et al., The Binding Database: data management and interface design. Bioinformatics, 2002. **18**(1): p. 130-9.
31. Chen, X., M. Liu, and M.K. Gilson, *BindingDB: a web-accessible molecular recognition database.* Comb Chem High Throughput Screen, 2001. **4**(8): p. 719-25.
32. Wang, R., et al., *The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures.* J Med Chem, 2004. **47**(12): p. 2977-80.
33. Wang, R., et al., *The PDBbind database: methodologies and updates.* J Med Chem, 2005. **48**(12): p. 4111-9.
34. Hu, L., et al., *Binding MOAD (Mother Of All Databases)*. Proteins, 2005. **60**(3): p. 333-40.
35. Smith, R.D., et al., *Exploring protein-ligand recognition with Binding MOAD.* J Mol Graph Model, 2006. **24**(6): p. 414-25.
36. Oprea, T.I., et al., *Lead-like, drug-like or "Pub-like": how different are they?* J Comput Aided Mol Des, 2007. **21**(1-3): p. 113-9.
37. Okuno, Y., et al., *GLIDA: GPCR--ligand database for chemical genomics drug discovery--database and tools update.* Nucleic Acids Res, 2008. **36**(Database issue): p. D907-12.
38. Sanders, M.P., et al., *ss-TEA: Entropy based identification of receptor specific ligand binding residues from a multiple sequence alignment of class A GPCRs.* BMC Bioinformatics, 2011. **12**(1): p. 332.
39. Li, J., et al., *Structure of bovine rhodopsin in a trigonal crystal form*. J Mol Biol, 2004. **343**(5): p. 1409-38.
40. Ballesteros, J.A.W., H., I*ntegrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein coupled receptors.* Methods Neurosci., 1995. **25**: p. 366-428.
41. Chinea, G., et al., *The use of position-specific rotamers in model building by homology*. Proteins, 1995. **23**(3): p. 415-21.
42. Lovell, S.C., et al., *The penultimate rotamer library.* Proteins, 2000. **40**(3): p. 389-408.
43. Madabushi, S., et al., *Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions*. J Biol Chem, 2004. **279**(9): p. 8126-32.
44. Oliveira, L., et al., *Sequence analysis reveals how G protein-coupled receptors transduce the signal to the G protein*. Proteins, 2003. **52**(4): p. 553-60.
45. Ye, K., et al., *A two-entropies analysis to identify functional positions in the transmembrane region of class A G protein-coupled receptors.* Proteins, 2006. **63**(4): p. 1018-30.
46. Bohm, B., et al., *The serine-protease inhibitor of cartilage matrix is not a chondrocytic gene product.* Eur J Biochem, 1992. **207**(2): p. 773-9.
47. Gunzer, F., et al., *Molecular detection of sorbitol-fermenting Escherichia coli O157 in patients with hemolytic-uremic syndrome*. J Clin Microbiol, 1992. **30**(7): p. 1807-10.
48. Bohm, H.J., *The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure.* J Comput Aided Mol Des, 1994. **8**(3): p. 243-56.
49. Klebe, G., *The use of composite crystal-field environments in molecular recognition and the de novo design of protein ligands*. J Mol Biol, 1994. 237(2): p. 212-35.
50. Rarey, M., et al., *A fast flexible docking method using an incremental construction algorithm.* J Mol Biol, 1996. **261**(3): p. 470-89.
51. Ye, K., et al., *Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting.* Bioinformatics, 2008. **24**(1): p. 18-25.

52. Barillari, C., G. Marcou, and D. Rognan, *Hot-spots-guided receptor-based pharmacophores (HS-Pharm): a knowledge-based approach to identify ligand-anchoring atoms in protein cavities and prioritize structure-based pharmacophores.* J Chem Inf Model, 2008. **48**(7): p. 1396-410.
53. Goodford, P.J., *A computational procedure for determining energetically favorable binding sites on biologically important macromolecules.* J Med Chem, 1985. **28**(7): p. 849-57.
54. Verdonk, M.L., J.C. Cole, and R. Taylor, *SuperStar: a knowledge-based approach for identifying interaction sites in proteins.* J Mol Biol, 1999. **289**(4): p. 1093-108.
55. de Graaf, C. and D. Rognan, *Selective structure-based virtual screening for full and partial agonists of the beta2 adrenergic receptor.* J Med Chem, 2008. **51**(16): p. 4978-85.
56. Chien, E.Y., et al., *Structure of the human dopamine D3 receptor in complex with a D2/D3 selective antagonist.* Science, 2010. **330**(6007): p. 1091-5.
57. Wu, B., et al., *Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists.* Science, 2010. **330**(6007): p. 1066-71.
58. Kufareva, I., *Status of GPCR modeling and docking in a community wide GPCR dock 2010 assessment*. Structure, 2011, **19**(8): p. 1108-26.
59. Vroling, B., et al., *GPCRDB: information system for G protein-coupled receptors.* Nucleic Acids Res, 2011, **39**(Database issue): p. D309-19
60. Leach, A.R., et al., *Three-dimensional pharmacophore methods in drug discovery*. J Med Chem. 2010, **53**(2): p. 539-58.
61. Okada, T., et al., *Functional role of internal water molecules in rhodopsin revealed by X-ray crystallography.* Proc Natl Acad Sci U S A, 2002. **99**(9): p. 5982-7.
62. Cherezov, V., et al., *High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor*. Science, 2007. **318**(5854): p. 1258-65.
63. Warne, T., et al., *Structure of a beta1-adrenergic G-protein-coupled receptor.* Nature, 2008. **454**(7203): p. 486-91.
64. Park, J.H., et al., *Crystal structure of the ligand-free G-protein-coupled receptor opsin.* Nature, 2008. **454**(7201): p. 183-7.
65. Hanson, M.A., et al., *A specific cholesterol binding site is established by the 2.8 A structure of the human beta2-adrenergic receptor.* Structure, 2008. **16**(6): p. 897-905.
66. Scheerer, P., et al., *Crystal structure of opsin in its G-protein-interacting conformation.* Nature, 2008. **455**(7212): p. 497-502.
67. Jaakola, V.P., et al., *The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist.* Science, 2008. **322**(5905): p. 1211-7.
68. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol, 1970. **48**(3): p. 443-53.
69. Krieger, E., G. Koraimann, and G. Vriend, *Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field.* Proteins, 2002. **47**(3): p. 393-402.
70. Vriend, G., *WHAT IF: a molecular modeling and drug design program.* J Mol Graph, 1990. **8**(1): p. 52-6, 29.
71. Delaunay, B., *Sur la sphère vide.* Izvestia Akademii Nauk SSSR, Otdelenie Matematicheskikh i Estestvennykh Nauk, 1934. **7**: p. 793-800.
72. Sanders, M., et al., *ss-TEA: Entropy based identification of receptor specific ligand binding residues from a multiple sequence alignment of class A GPCRs.* BMC Bioinformatics, 2011. **12**: p. 332
73. Kumar, S. and R. Nussinov, *Relationship between ion pair geometries and electrostatic strengths in proteins.* Biophys J, 2002. **83**(3): p. 1595-612.
74. Renner, S. and G. Schneider, *Fuzzy pharmacophore models from molecular alignments for correlation-vector-based virtual screening.* J Med Chem, 2004. **47**(19): p. 4653-64.
75. Breu, V., et al., *Separable binding sites for the natural agonist endothelin-1 and the non-peptide antagonist bosentan on human endothelin-A receptors.* Eur J Biochem, 1995. **231**(1): p. 266-70.
76. Shi, L. and J.A. Javitch, *The binding site of aminergic G protein-coupled receptors: the transmembrane segments and second extracellular loop.* Annu Rev Pharmacol Toxicol, 2002. **42**: p. 437-67.
77. Kane, B.E., B. Svensson, and D.M. Ferguson, *Molecular recognition of opioid receptor ligands.* AAPS J, 2006. **8**(1): p. E126-37.
78. Li, J.G., et al., *ASP147 in the third transmembrane helix of the rat mu opioid receptor forms ion-pairing with morphine and naltrexone*. Life Sci, 1999. **65**(2): p. 175-85.

79. Surratt, C.K., et al., -mu opiate receptor. *Charged transmembrane domain amino acids are critical for agonist recognition and intrinsic activity.* J Biol Chem, 1994. **269**(32): p. 20548-53.
80. Sadowski, J., J. Gasteiger, and G. Klebe, *Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures*. J Chem Inf Comput Sci, 1994. **34**(4): p. 8.
81. Liu, X., et al., *Cyndi: a multi-objective evolution algorithm based method for bioactive molecular conformational generation*. BMC Bioinformatics, 2009. **10**: p. 101.
82. PipelinePilot, Pipeline Pilot, Scitegic, Inc: San Diego.
83. Canutescu, A.A., A.A. Shelenkov, and R.L. Dunbrack, Jr., *A graph-theory algorithm for rapid protein side-chain prediction.* Protein Sci, 2003. **12**(9): p. 2001-14.
84. Warne, T., et al., *The structural basis for agonist and partial agonist action on a beta(1)-adrenergic receptor.* Nature, 2011. **469**(7329): p. 241-4.
85. Krieger, E., et al., *Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8.* Proteins, 2009. **77**(Suppl 9): p. 114-22.
86. Clark, R.D. and D.J. Webster-Clark, *Managing bias in ROC curves*. J Comput Aided Mol Des, 2008. **22**(3-4): p. 141-6.
87. Jain, A.N. and A. Nicholls, *Recommendations for evaluation of computational methods.* J Comput Aided Mol Des, 2008. **22**(3-4): p. 133-9.
88. Landrum, G., RDKit (www.rdkit.org).
89. Staubert, C., et al., *Structural and functional evolution of the trace amine-associated receptors TAAR3, TAAR4 and TAAR5 in primates.* PLoS One. **5**(6): p. e11133.
90. Rosenbaum, D.M., et al., *GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function.* Science, 2007. **318**(5854): p. 1266-73.
91. Holst, B., et al., *Overlapping binding site for the endogenous agonist, small-molecule agonists, and ago-allosteric modulators on the ghrelin receptor.* Mol Pharmacol, 2009. **75**(1): p. 44-59.
92. Macdonald, D., et al., *Molecular characterization of the melanin-concentrating hormone/receptor complex: identification of critical residues involved in binding and activation.* Mol Pharmacol, 2000. **58**(1): p. 217-25.
93. Fierens, F.L., et al., *Lys(199) mutation of the human angiotensin type 1 receptor differentially affects the binding of surmountable and insurmountable non-peptide antagonists.* J Renin Angiotensin Aldosterone Syst, 2000. **1**(3): p. 283-8.
94. Khasawneh, F.T., et al., *Differential mapping of the amino acids mediating agonist and antagonist coordination with the human thromboxane A2 receptor protein.* J Biol Chem, 2006. **281**(37): p. 26951-65.
95. Xu, F., et al., *Structure of an agonist-bound human A2A adenosine receptor.* Science, 2011. **332**(6027): p. 322-7.
96. Muegge, I. and S. Oloff, *Advances in virtual screening.* Drug Discov Today: Technologies, 2006. **3**(4): p. 405-411.
97. Ortore, G., et al., *Different Binding Modes of Structurally Diverse Ligands for Human D3DAR.* J Chem Inf Model, 2010. **50**(12): p. 2162-75

CHAPTER 6

In silico veritas: the pitfalls and challenges of predicting GPCR-ligand interactions

Luc Roumen[1], Marijn P.A. Sanders[2], Bas Vroling[2], Iwan J.P. de Esch[1], Jacob de Vlieg[2], Rob Leurs[1], Jan P.G. Klomp[3], Sander B. Nabuurs[2] and Chris de Graaf[1]

[1]Department of Medicinal Chemistry, VU University Amsterdam, Amsterdam, the Netherlands; [2]CMBI, NCMLS, Radboud University Nijmegen Medical Centre, Nijmegen, the Netherlands; [3]Department of Molecular Design and Informatics, MRL, MSD, Oss, the Netherlands

Pharmaceuticals 2011, 4(9), 1196-1215

Acknowledgments

We would like to thank Gert Vriend and Martine Smit for their contributions and helpful discussions. This work is supported by the Top Institute Pharma [project number D1.105: the GPCR Forum]. S.N. is supported by the Netherlands Organization for Scientific Research (NWO) through a VENI grant (700.58.410). C.de.G. is supported by the Netherlands Organization for Scientific Research (NWO) through a VENI grant 700.59.408).

Abstract

Recently the first community-wide assessments of the prediction of the structures of complexes between proteins and small molecule ligands have been reported in the so-called GPCR Dock 2008 and 2010 assessments. In the current review we discuss the different steps along the protein-ligand modeling workflow by critically analyzing the modeling strategies we used to predict the structures of protein-ligand complexes we submitted to the recent GPCR Dock 2010 challenge. These representative test cases, focusing on the pharmaceutically relevant G Protein-Coupled Receptors, are used to demonstrate the strengths and challenges of the different modeling methods. Our analysis indicates that the proper performance of the sequence alignment, introduction of structural adjustments guided by experimental data, and the usage of experimental data to identify protein-ligand interactions are critical steps in the protein-ligand modeling protocol.

6.1. Introduction

In the last few years, the disciplines involved in *in silico* protein structure prediction have greatly evolved. Not only have more tools and modeling programs become available, but also the amount of varying approaches to produce predictive models has increased. Structure prediction of protein-ligand complexes by comparative or homology modeling can be subdivided into the following major steps: (1) identification of homologue proteins for which a three-dimensional structure is available; (2) alignment of the target sequence with the sequence of the template structure; (3) building the coordinates of the three-dimensional model of the target; (4) modeling the protein-ligand interactions; and (5) assessing ligand binding mode prediction accuracy by investigating ligand structure activity data or biological data [1, 2].

Critical assessments of *in silico* methods to predict the structure of proteins (CASP [3]) and protein-protein complexes (CAPRI [4]) have been established in the past years, and many comparative docking studies to predict the binding orientation of small molecule ligands in known protein structures have been reported [5]. Only very recently, however, the first community-wide assessments of the prediction of the structures of complexes between proteins and small molecule ligands have been reported in the so-called GPCR Dock assessments [6, 7]. The first was initiated in 2008 to predict conformation of the human adenosine A2A receptor in complex with the small ligand ZM241385 [7, 8]. The second was organized in 2010 [6] to predict the conformation of the dopamine D3 receptor in complex with the small ligand eticlopride [9], as well as the chemokine receptor CXCR4 bound to the small ligand 1t [10] or the cyclic peptide CVX15 [6, 10]. These assessments did not only give the protein modeling community the chance to objectively (and prospectively) test their methods to predict the structure of complexes between proteins and small (drug-like) ligands, but also offered a unique opportunity to identify the problems and pitfalls in the prediction of protein-ligand interactions. In the current review we will discuss the different steps along the protein-ligand modeling workflow by critically analyzing the modeling strategies we used to generate the structures we submitted to the GPCR Dock 2010 challenge. These representative test cases will be used to demonstrate the strengths and challenges of the different methodologies and their impact on modeling accuracy.

6.2. Experimental Section

6.2.1. GPCR Dock 2010

In spring 2010, the group of Stevens et al. challenged the scientific community to participate in the structure prediction assessment GPCR Dock 2010. The subject of the challenge consisted of three different crystal structures for which multiple models could be deposited. The first case encompassed modeling the dopamine D3 receptor co-crystallized with the antagonist eticlopride [9] (**Figure 6.1**). The dopamine D3 receptor is closely related to the adrenergic beta 1 and 2 receptors, for which a crystal structure

has already been elucidated [11, 12]. The aminergic receptor family possesses a high sequence identity for the residues involved in ligand binding, including D3.32, S5.43, S5.46, and Y7.43 [13-16]. Due to the functional similarity and identical binding sites of the target to the adrenergic receptors, it was considered as the easiest of the three challenges.

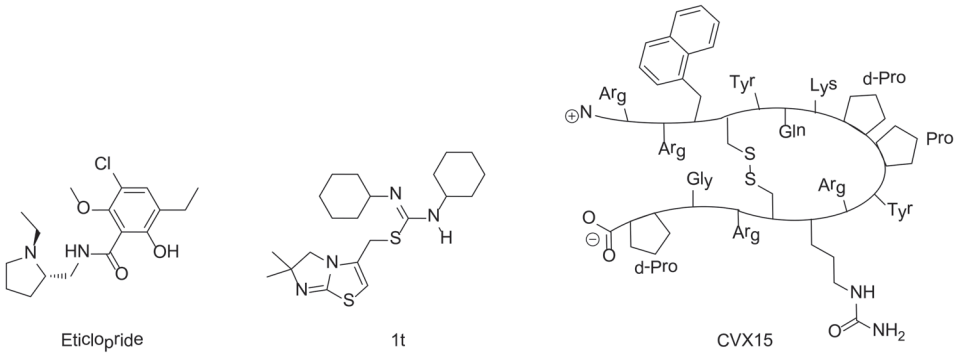


Figure 6.1: Chemical structures of eticlopride, 1t and CVX15. The first is co-crystallized with the dopamine D3 Receptor and the latter two with the chemokine Receptor CXCR4.

The second case encompassed constructing a model for the chemokine receptor CXCR4 co-crystallized with a small ligand 1T [17] (**Figure 6.1**). The similarity of the chemokine receptor CXCR4 in sequence and function is distant compared to the GPCRs for which a crystal structure has been elucidated [10] and as such was expected to pose a more difficult challenge than the dopamine D3 receptor complex. The importance of certain amino acids for the binding of small ligands has been described [18-20], and the ligand in question is one from a compound series for which structure activity relationship data has been determined [17]. These data could be used to derive binding hypotheses for the ligand in the CXCR4 model.

The third case involved building a model for the chemokine receptor CXCR4 co-crystallized with a large cyclic peptide, the antagonist CVX15 [21, 22] (**Figure 6.1**). Due to the large conformational space accessible to the ligand, it was considered the most difficult challenge in GPCR Dock 2010: none of the deposited models were able to predict any of the critical contacts between protein and ligand [6]. As the current review focuses on the prediction of interactions between proteins and small (drug-like) ligands, the CVX15 modeling case will not be described in this review.

6.2.2. Modeling Approaches

Our two research groups have taken different approaches to construct the structural models for the challenge. The CDD-CMBI group contested in both small ligand challenges and ranked second for the dopamine D3 receptor and fourteenth for the chemokine CXCR4 receptor, whereas the VU-MedChem group only contested in the CXCR4-1t challenge for which they obtained first rank [6]. The CDD-CMBI group chose to combine

sequence conservation knowledge with the placement of a pharmacophore definition as a description for the protein-ligand interactions in space (**Figure 6.2A**). In parallel, a homology model was constructed based on a template structure, which was subsequently combined with the pharmacophore and flexible receptor docking to obtain a structural model for the protein-ligand interactions. The VU-MedChem group constructed a structural model after which literature data was used to determine the most likely protein-ligand interactions (**Figure 6.2B**). Each of the steps taken by our groups is graphically outlined below.

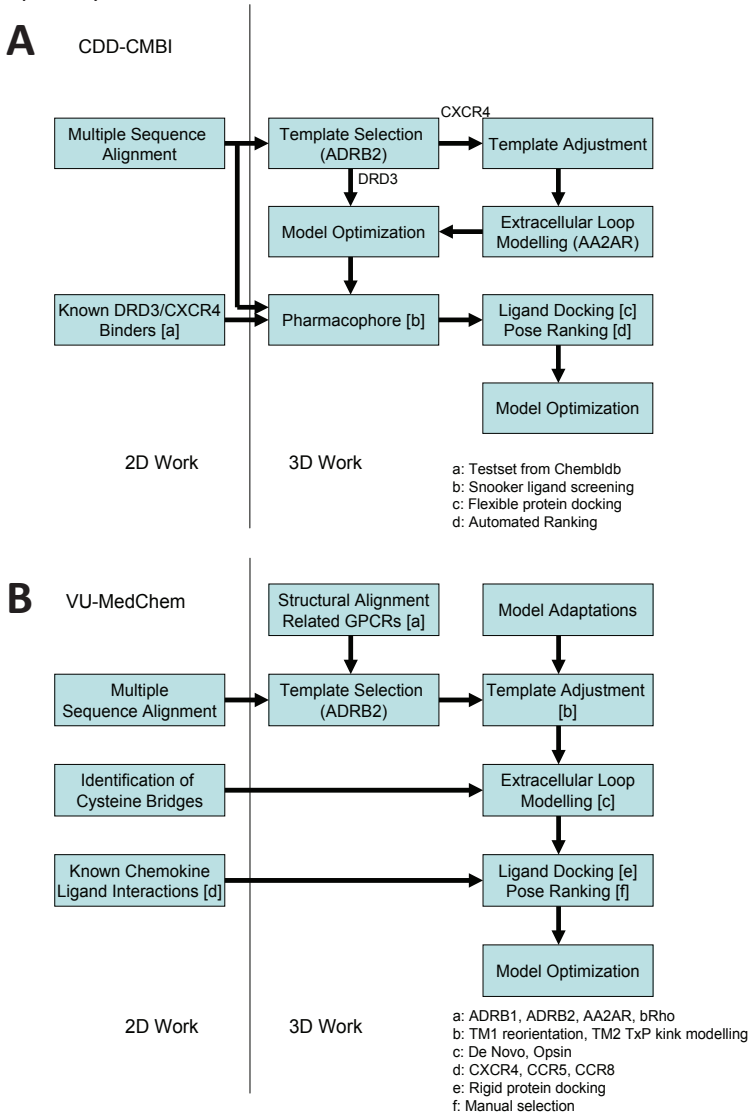


Figure 6.2A: Modeling workflow as used by the CDD-CMBI group in the GPCR Dock 2010 challenge. **B:** Modelling workflow as used by the VU -MedChem group in the GPCR Dock 2010 challenge.

6.2.3. Sequence Analysis

6.2.3.1. Current Approaches

The construction of a multiple sequence alignment of the target sequence with a series of potential homologues is an important step in predicting the structure of the target protein [2]. Potential misalignments have a direct impact on the location of amino acids in the protein model, which in turn directly influences the models ability to correctly predict protein-ligand interactions. As such, the accuracy of the alignment should be optimized as much as possible. To accomplish this, common practice is to include sequences belonging to target family members in an attempt to identify important residues and sequence motifs that might point to a similarity in protein fold or function. Highly conserved amino acids indicate the conservation of the general protein function, three-dimensional fold and structural features. Amino acids that are very different between homologous proteins potentially indicate locations of specificity to binding partners such as small ligands or other proteins.

A G protein-coupled receptor contains seven transmembrane helices, each of which contains highly conserved residues and sequence motifs (DRY in TM3, CWxP in TM6 and NPxxY in TM7) likely related to generic receptor activation. Based on sequence and motif conservation, Ballesteros and Weinstein [23] developed a generic numbering scheme for GPCRs. This numbering scheme allows for consistent residue numbering across multiple proteins, independent of their sequential numbers. The underlying principle is that residues with the same general residue number have equivalent locations in their tertiary structures and consequently in the multiple sequence alignments. Residue numbers are in the format 'X.Y', where X indicates the TM helix, and Y the residue position with respect to the most conserved residue position in the helix, which gets the number 50. In addition to these conserved residues, most class A GPCRs also contain conserved cysteine residues in both TM3 (C3.25) and the extracellular loop 2 (C45.50) that together form a cysteine bridge. Resulting from these observations, the alignment of the structural aspects for GPCRs is most challenging for residues arranged outside of the transmembrane helices, in the N-terminus, intracellular and extracellular loops, and the C-terminus. However, most GPCR ligands are interact with the transmembrane domain [24], making the construction of a protein-ligand binding model feasible.

6.2.3.2. DRD3 Case

Of the crystallized GPCRs the adrenergic beta 2 receptor (ADRB2) has the highest sequence similarity with the dopamine D3 receptor, making this the most suitable template for modeling studies. The alignment of the dopamine D3 sequence with the ADRB2 sequence was based on alignments from the GPCRDB [25], with additional manual refinements in the loop regions. A sequence-based prediction of residues involved in ligand interaction was made for all transmembrane residues. Residues were scored according to ligand binding probability based on an analysis of Shannon entropies of residue positions

[26] of a multiple sequence alignment of around 7700 class A GPCR transmembrane domains. The alignment included 64 dopamine receptor D2 and D3 sequences of in total 34 species. The most important ligand interacting residues were predicted to be D3.32, V3.33, S5.42, H6.55, Y7.35 and T7.39, which is corroborated by mutation data [14-16].

6.2.3.3. CXCR4 Case

In case of the CXCR4 challenge, the multiple sequence alignment study was supplemented with a literature study in order to investigate the target for conservation, specific family motifs and the importance of amino acids involved with ligand binding. A sequence alignment of all CXC chemokine receptors was constructed by sequence retrieval from the Uniprot database [27] and alignment with ClustalW [28]. The likelihood of cysteine bridges was assessed based on their sequence conservation (**Figure 6.3**). Similar to all other class A GPCRs, CXCR4 contains the cysteine residues C3.25 and C45.50. However, the sequence analysis also showed that all CXCR isoforms except for CXCR6 contain a cysteine residue both in the N-terminus as well as the extracellular loop 3. From these observations, it was concluded that for CXCR4 an additional cysteine bridge should be incorporated into the protein model between the N-terminus and ECL3. The presence of both cysteine bridges were indeed confirmed by the CXCR4 crystal structures [10].

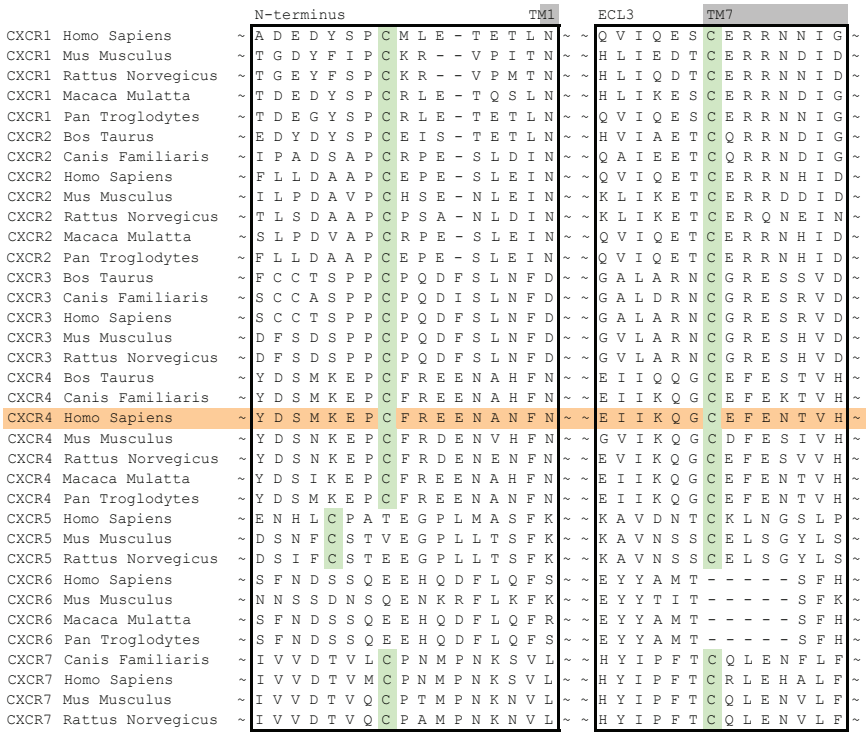


Figure 6.3: Sequence alignment of the CXCR family including various species. Indicated are the conserved cysteine residues in the N-terminus and the extracellular loop 3 which are hypothesized to form a cysteine bridge.

6.2.4. Template Selection and Construction

6.2.4.1. Current Approaches

Template selection is mostly based on the sequence similarity between the template structure and the target sequence. If the sequence identity is low (<20%) and the structural aspects of the model are not highly conserved (alpha-helices and beta-sheets), it is very challenging to construct a predictive model for the protein fold [2] and residues interacting with potential ligands. For GPCRs, the crystal structures available for various proteins [8, 11, 12, 29] have shown only slight differences in the general arrangement of the seven transmembrane helices. The only marked differences were observed when comparing the activated [29] and inactivated [30] conformations of the rhodopsin structure. Here, a reorientation of amino acids in the central and intracellular regions of the TM region as well as a twisting of TM5 and TM6 are observed. Currently, novel insights in the conformational changes upon activation can also be derived from the activated AA2AR [31] and ADRB2 [32, 33] structures.

Due to the structural conservation, model construction based on any of the crystal structures would likely result in a preliminary model of the transmembrane region with most of the amino acids correctly pointing into the helical bundle [1, 24]. Minor adjustments can be made to the alpha helices in case gaps, insertions or helical kinks are expected. This holds particularly true for the occurrence of proline residues in helices combined with either serine or threonine TM residues [34, 35]. Adjustments of the template to accommodate these differences can be accomplished by constructing a custom template for the helix and using molecular dynamics or Monte Carlo simulations to predict the overall helical fold.

With regard to the loops, there are noticeable differences, especially in the second extracellular loop, which seems to show a different fold in all currently known protein families. In bovine rhodopsin this loop folds into a beta sheet structure, the adrenergic beta 1 and 2 receptors exhibit an alpha helical fold, and the adenosine A2A receptor shows a coil structure. Using experimental mutation data on the extracellular loop residues, one can produce a protein-ligand interaction model and restrain amino acids from the loop in the binding site while optimizing the rest of the loop structure [36-39]. However, when no data is available, it is advised to omit the modeling of the extracellular loops altogether and prioritize the modeling of interactions with residues in the TM bundle. Multiple template structures can be used as a basis for the construction of the final model, such as a combination of a template for the transmembrane domain and a template for the loop regions.

6.2.4.2. DRD3 Case

Since it has the highest similarity to the dopamine D3 receptor, the ADRB2 structure [11] (PDB code 2RH1) was chosen as the modeling template. Since the residues in the loop between TM5 and TM6 (residues 218-317) were not present in the crystal structure due to the insertion of T4 lysozyme, these residues were discarded in the modeling process. No additional modifications to the structure were deemed necessary based on the sequence.

6.2.4.3. CXCR4 Case

A structural assessment of available crystal structures of other G protein-coupled receptors was ensued following the alignment of their sequences to those of the chemokine receptor family. The structural alignment was performed with MOE2009.10 [40] and consisted of the structures of the human adenosine A2A Receptor [8] (PDB: 3EML), human adrenergic Beta 1 Receptor [12] (PDB: 2VT4), human Adrenergic Beta 2 receptor [11] (PDB: 2RH1) and the bovine opsin structure [29] (PDB: 3DQB). The selection and construction of a template structure posed several challenges. Overall, none of the sequences of the crystal structures possessed a significantly higher sequence similarity with the CXCR4 sequence than another, and as such, the choice of a template structure seemed arbitrary. In the end, the crystal structure of ADRB2 was found to be a feasible template due to the importance of D3.32, N7.39 and Y7.43 in the binding of the antagonist carazolol [11], and the analogous involvement of residues at these positions in chemokine receptor binding by small ligands [20, 41-43]. From the sequence analysis and literature [44] it became clear that the chemokine receptor family possessed a unique TxP motif in TM2 which when aligned to the available crystal structures would produce a gap at the top of the transmembrane helix (**Figure 6.4**). A misalignment of this region would not put important amino acids into the TM bundle such as D2.63 [20]. To overcome this problem, we decided to customize the conformation of TM2 to accommodate the difference in amino acid rather than using an existing crystal structure as a template. In a manner similar to Govaerts et al. [44], a three-dimensional model of the helix was constructed to predict the helical bend of the TxP motif using the AMBER program [45] which was subsequently incorporated into the template. Lastly, TM1 of the template structure was oriented closer to TM7 since there is a spatial gap between these helices in the crystal structure due to the uncapped N-terminus of TM1. It was reasoned that the smaller residue at position 7.40 in CXCR4 compared to that of in the ADRB2 structure (A vs. W, respectively) could accommodate the helical repositioning.

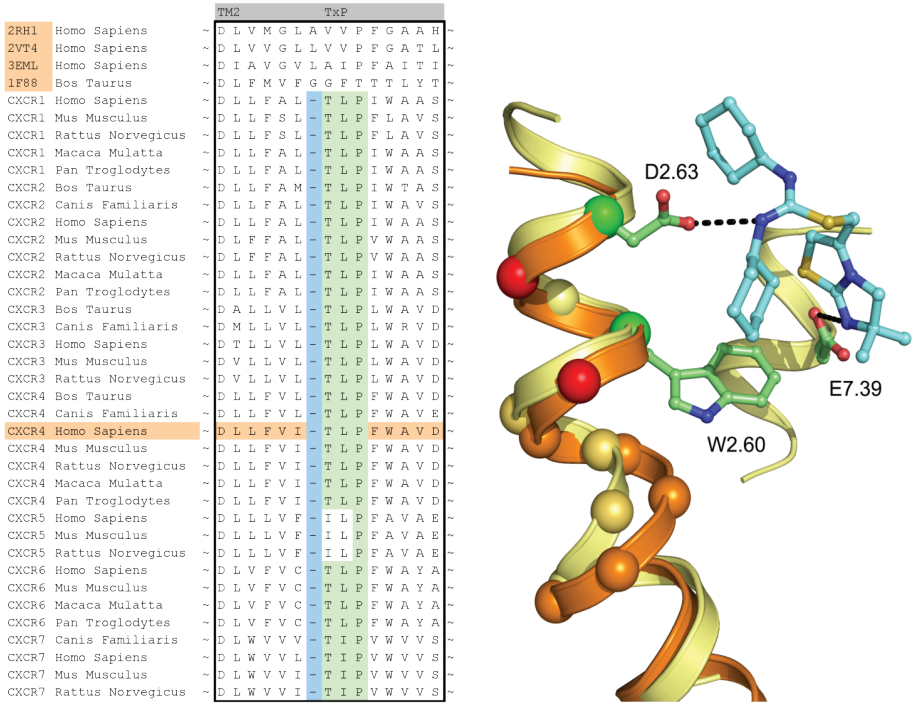


Figure 6.4: Sequence conservation of the TxP motif in the CXCR family and the implication of a possible misalignment on the amino acids pointing into the TM bundle. Using the unadjusted TM2 template of ADRB2 (orange ribbon and Cα spheres) would place W2.60 and D2.63 on the red Cα spheres, where the residues would not contact the ligand. In the CXCR4 crystal structure (yellow ribbon and Cα spheres), W2.60 and D2.63 are positioned on the green Cα spheres, indicating a shift from the ADRB2 template.

6.2.5. Homology Model Construction

6.2.5.1. Current Approaches

There are many homology modeling software packages available that allow the user to construct a three-dimensional model for their desired target sequence based on a template structure. Most packages work in a similar fashion. Residues that are in common between the template and the target are typically kept unaltered in the initial step of the model construction. Next, residues that differ are mutated and placed in an initial conformation based on a rotamer library [46, 47] that consists of the most commonly observed orientations of the residues throughout various crystal structures in the Protein Data Bank [48]. Residues are perturbed and reoriented whilst decreasing the amount of steric clashes and increasing the amount of stabilizing interactions. Finally, multiple different homology models are created and if desired optimized by the modeling program. It is recommended to investigate the predictive value of the model by identifying potential protein-ligand interactions as hypothesized by the user or as determined from literature data, as described below.

6.2.5.2. DRD3 Case

The template was prepared by a cleanup of the structure (removal of waters, sulfate ions, maltose, acetamide and butanediol), and removal of the T4 lysozyme protein. The lipids were retained for modeling. Using the alignment and the template as the starting point, modeling was performed with the Yasara program and its built-in modeling algorithm [49]. Side chains were added with Yasaras implementation of SCWRL [50], and then the model was subjected to an energy minimization with the Yasara2 force field as described previously [49]. WHAT CHECK [51] validation scores were used to score and rank the final models. The generated model was refined manually and a final energy minimization step was applied to relax the atoms of the DRD3 model.

6.2.5.3. CXCR4 Case

Structural modeling of the CXCR4 structure was commenced with the TM bundle and most of the loop structures. Three structural regions were excluded from the initial model and were constructed later. These included the extracellular loop 2 (ECL2), the N-terminus, and the C-terminus. The extracellular loop 2 could not be modeled using any of the protein crystal structures available since the length of the CXCR4 loop is different and amino acids would either have to be added to or deleted from the (template) sequence. Different conformations of ECL2 were generated using MODELLER [52] and incorporated into the CXCR4 models, and one conformation was constructed based on the ECL2 of the opsin structure [29] (PDB: 3DQB). A structural model for the N-terminus could be retrieved from the crystal structure of the ligand CXCL12 [53] (PDB: 2K04) after positioning the arrangement on top of the TM bundle of the CXCR4 model. The C-terminus was modeled with a random arrangement since all atoms were requested for submission in the GPCR Dock assessment.

6.2.6. Ligand Interaction Modeling

6.2.6.1. Current Approaches

The prediction of protein-ligand interactions can be a challenging task for which the availability of experimental data is often beneficial. Many different experimental methods have been applied to identify GPCR ligand binding sites, and elucidate protein conformations and protein-ligand interactions, including site-directed mutagenesis studies [54], infrared probes [55], NMR spectroscopy [56], fluorescence measurements [57, 58], and the use of unnatural amino acids [59] and amino acid chelators [43]. Each of these tools provides information about amino acids interacting either with each other or with the ligand. In addition to these pharmacological and biophysical methods, a common method is the evaluation of other receptor binders and structural analogues of the ligand with regard to their interaction with the protein [17, 60] (and protein mutants if available). The majority of the methods allow the identification of residues interacting with the ligand thereby providing an anchor for the arrangement of

the ligand in the binding site. Thus model validation can be performed by investigating both structure activity relationships based on ligand analogues, but it can also utilize site-directed mutagenesis data.

Molecular docking methods can be used to generate an initial binding pose of the small molecule ligand in the protein model. Docking programs rank the poses based on a scoring function that is knowledge-based, energy-based or empirical [5]. Since the scoring functions are optimized for protein-ligand complexes such as in the Protein Data Bank, one may assume that the results are precise. However, when little is known about the protein binding cavity or the rotamer conformations of the amino acids in the binding pocket, one cannot be certain that the docking program will generate or select the correct binding mode. In such case, it is helpful to post-process the binding poses using filtering schemes that prioritize specific residue interactions [1].

6.2.6.2. DRD3 Case

Ligand interactions were modeled using pharmacophore searches followed by flexible receptor docking. Structure-based pharmacophores derived from the transmembrane domains of the constructed homology models were built using the Snooker program [61]. In short, the Snooker method scores residue positions based on sequence conservation in a multiple sequence alignment, and deduces the key interacting residues from the derived statistics. Known DRD3 actives were retrieved from Chembl [62] using a 50nM activity cutoff. Conformations of the ligands were generated using Cyndi [63]. A pharmacophore search was performed to identify the structure-based pharmacophore complementary to most of the active compounds. The donor and positive ionizable features in the resulting pharmacophore originate from D3.32, the acceptor feature from T7.39, and the hydrophobic features from F6.51, F6.52, H6.55 and F7.35. Subsequently, eticlopride was matched in this structure-based pharmacophore (**Figure 6.5A,B**). The low-resolution binding modes obtained from Snooker were used to guide high-resolution molecular docking by the Fleksy program [64]. The various orientations of eticlopride in the DRD3 receptor model were used as anchors to guide induced fit docking using the Fleksy protocol. The generated poses were optimized in the homology model (**Figure 6.5C)** and ranked using a consensus scoring function which utilizes docking scores, geometrical quality indicators and molecular dynamics force field interaction energies. All final poses contained the charged interaction of the basic amine with D3.32 and the hydrophobic interactions with TM3, TM5, TM6 and TM7 similarly to carazolol in the ADRB2 receptor complex and timolol in the ADRB1 receptor complex. A retrospective comparison to the DRD3 crystal structure revealed that the polar interaction with D3.32 was correctly predicted, as well as the hydrophobic contacts to V3.33, V5.39, V5.39, H6.55, F7.35, T7.39 and Y7.43 (**Figure 6.5C,D**). Many of these residues are in agreement with the pharmacophore features defined by the Snooker protocol.

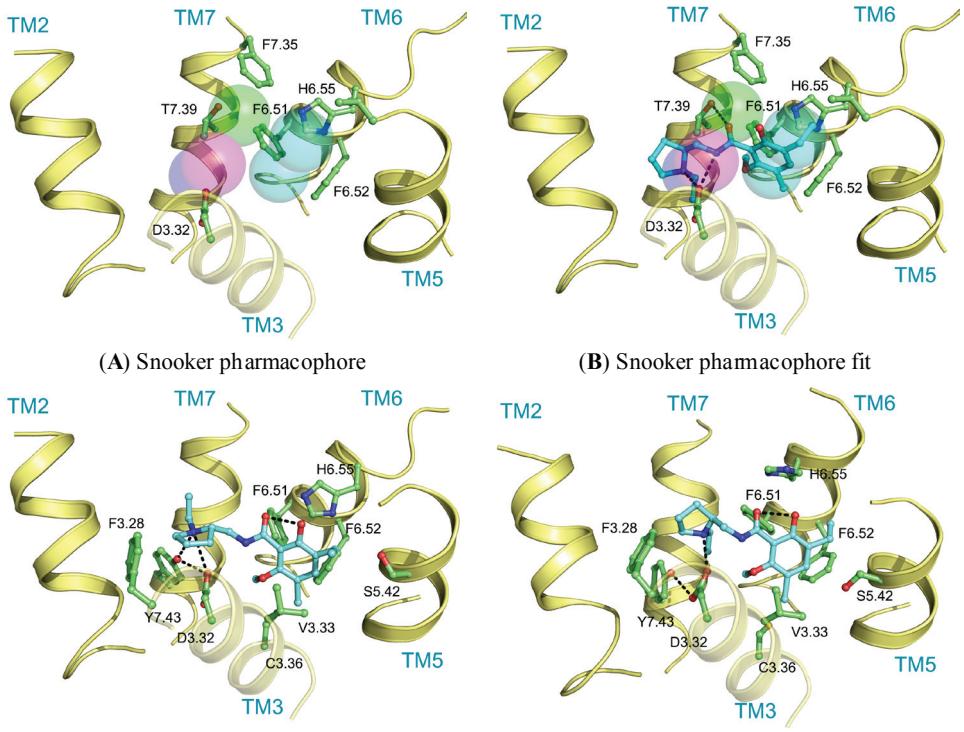


Figure 6.5: Binding pose as predicted by CDD-CMBI using Snooker and Fleksy compared to the X-ray structure.

The final GPCR Dock 2010 assessment of the five submitted models placed each of them in the top 10 of all 117 submitted models with the best model ranking second. Despite a relatively high receptor model RMSD the submitted models were able to capture a large part (35% to 57%) of the receptor-ligand atomic contacts, including the hydrogen bond with D3.32 and the hydrophobic interactions with F6.51, F6.52, H6.55 and F7.35. The best binding mode was able to capture 36 of 65 atomic contacts as well as 12 of 15 residues directly interacting with the ligand. Interestingly, the accuracy of the five submitted DR3D models is in excellent agreement with the ranking generated by the Fleksy consensus scoring function [6] (**Figure 6.6**). This highlights the potential of knowledge based scoring functions in the identification of near-native receptor-ligand complex geometries. The scoring function performs less well in ranking the much less accurate solutions generated by CDD-CMBI for the CXCR4 target (**Figure 6.6**), but does correctly assign them a worse score compared to the DRD3 solutions.

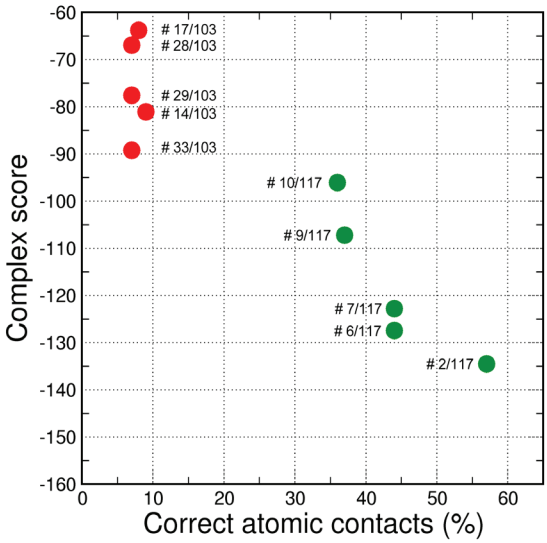
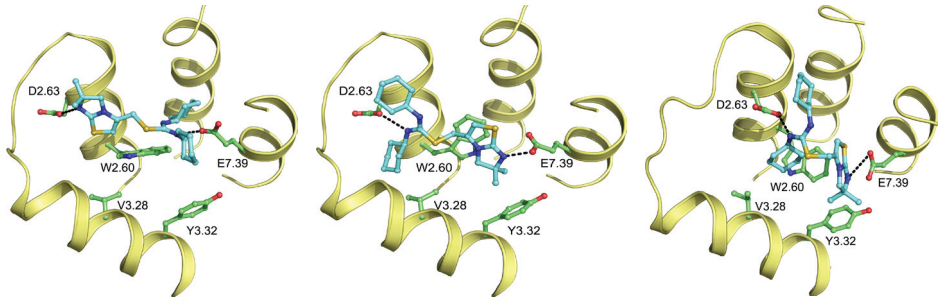


Figure 6.6: Ranking of the submitted models by the Fleksy consensus score. The five submitted models for DRD3 are shown in green, for comparison the five submitted models by CDD-CMBI for CXCR4 are shown in red. The consensus complex score (lower is better) is plotted against the percentage of correct contacts as determined in the final GPCR dock 2010 assessment. The final rank in the GPCR dock 2010 assessment is indicated for each of the solutions.

6. 2.6.3. CXCR4 Case

Wong [20] and Rosenkilde [43] determined that negatively the charged residues D2.63, D4.60, D6.58 and E7.39 play a role in the CXCR4 binding of several antagonists. Based on the physico-chemical properties of ligand 1T (**Figure 6.1**), it seemed likely protonated on one of the thiourea moieties. As such, the negatively charged residues were the first candidates for an anchor point of the ligand in the TM bundle of CXCR4. Eight possible binding poses were conceived in which the ligand is sandwiched between two negatively charged residues that interact with either of the thiourea moieties (D2.63 was only able to combine with E7.39). From these poses, the five poses that corroborated the structure-activity relationships as described by Thoma et al. [17], were chosen for the final models. The ligands were docked into the models using GOLD [65] using the restraints of ionic interactions. Both binding modes interacting with D2.63 and E7.39 were among those chosen for submission to GPCR Dock 2010 (**Figure 6.7A,B**). The final models were optimized using AMBER molecular dynamics simulation including the ionic interactions, followed by an unconstrained energy minimization and molecular dynamics to optimize the overall structure. Model ranking was performed by visual inspection.

Compared to the crystal structure, the 5th ranked model possessed the most similar binding mode in the CXCR4 structure. In total, 19 of 64 atomic contacts from five of 13 residues were correctly predicted by the model. Although this might seem a low amount, the model did capture the hydrophilic interactions of the ligand with D2.63 and E7.39 (**Figure 6.7B,C**), and the fraction of the pocket that was predicted correctly reached 45% [6]. The major discrepancy between model and crystal structure is the folding of ECL1 and ECL2 to place the correct residues in the binding cavity.



(A) VU alternate pose (B) VU most accurate pose (C) Xray pose

Figure 6.7: Binding poses as predicted by VU-MedChem using GOLD compared to the X-ray structure.

The other research groups that participated in the GPCR Dock 2010 were able to reproduce the interaction of 1t with either D2.63 or E7.39, but were unable to produce a pose that interacted with both ionic residues. The explanation for this observation can be twofold. Firstly, the location of the ligand 1t in the transmembrane domain is new, namely, inside the minor pocket [24], whereas all potential template structures portrayed their ligand in the major pocket. Secondly, the correct spatial construction of the binding cavity is dependent on the sequence alignment and resolving the influence of the TxP motif on TM2, as well as the positioning of TM1 due to the presence of the cysteine bridge between the N-terminus and ECL3 [10]. The VU-MedChem group was able to capture the correct interactions in the binding cavity because they focused their attention on these particular regions.

6.2.6.4. Water Molecules

Although none of the modeling attempts have included the prediction of protein-water or ligand-water interactions, it should be noted that conserved water clusters have been identified for class A GPCRs between TM1, TM2, TM6 and TM7. These conserved waters are suggested to be involved in receptor activation [66] and can in principle be included in GPCR modeling procedures [67]. However, none of these water molecules is directly involved in water-mediated protein-ligand interactions. In the AA2AR crystal structures, water molecules are included in protein-ligand hydrogen bonding networks [8] and consideration of some of these water

molecules has been shown to improve structure-based virtual screening accuracy [68]. In the ADRB1, ADRB2 and DRD3 crystal structures, no water molecules have been resolved in the vicinity of the ligand, and none of the crystallographic water molecules in contact distance of the ligand in the CXCR4 structure mediate polar protein-ligand interactions. Moreover, given the hydrophobicity of the CXCR4 pocket, it is unclear to which extent the water molecules influence ligand binding; hence it would have been difficult to predict the position of the water molecules in the pocket. Finally, the need to include water molecules in the prediction of protein-ligand interactions is target-dependent, as demonstrated by comparative docking studies [69, 70].

6.3. Conclusions

Structure-based modeling and design can aid in understanding and optimizing protein-ligand interactions and as such has proven a valuable tool in modern drug discovery. The approaches to structure-based modeling have evolved to include prior knowledge, which greatly aids the identification of ligand binding cavities as well as the validation of generated ligand binding poses. In the GPCR Dock 2010 challenge, our groups used different approaches to obtain predictions for the protein-ligand interactions of eticlopride and 1t in the DRD3 and CXCR4 crystal structure, respectively.

The largely automated use of extensive sequence analysis in the derivation of the structure-based pharmacophores followed by the selection of the structure-based pharmacophore best correlating to a large number of known actives, resulted in a pharmacophore definition which correctly encoded the crucial protein ligand interactions for DRD3. The simultaneous optimization of protein-ligand interactions and the protein and ligand structure themselves in a restraint docking procedure allowed that correct atomic contacts were produced and that these predicted contacts were optimized. Importantly, the applied knowledge based scoring function was able to correctly identify and assign the highest rank to the best near-native complex geometry without manual intervention.

Resulting from a knowledge-based investigation of the CXCR sequence, the CXCR4 model required a template adjustment in TM1 and TM2 to accommodate an expected cysteine bridge in the N-terminus as well as a kink in TM2. Inclusion of such detail resulted in a highly customized CXCR4 model which was able to correctly capture the most important protein-ligand interactions as observed in the crystal structure. Without in depth sequence analysis, experimental receptor knowledge and knowledge on ligand analogues to validate and refine the ligand binding prediction, the result would not have been as successful.

From this study we conclude that the integration of experimental target data and in silico studies in iterative cycles is of prime importance for the accurate prediction of the protein architecture as well as ligand binding modes therein. Even then, the elucidation of a ligand binding mode is not always clear cut, since symmetry

in the ligand (such as for 1t) or binding pocket increases modeling difficulty. In the future, the inclusion of experimental data and further development of more automated procedures will help to fill the gaps that still exist in the GPCR structural landscape, even when GPCRs belonging to different subfamilies are being resolved by crystallography in rapid succession (http://gpcr.scripps.edu/) [31-33, 71].

References

1. de Graaf, C. and D. Rognan, *Customizing G Protein-coupled receptor models for structure-based virtual screening.* Curr Pharm Des, 2009. **15**(35): p. 4025-48.
2. Eswar, N., et al., *Comparative Modeling of Drug Target Proteins*, in *Comprehensive Medicinal Chemistry II*. 2007, Elsevier: Oxford. p. 215-36.
3. Moult, J., et al., *A large-scale experiment to assess protein structure prediction methods.* Proteins: Structure, Function, and Bioinformatics, 1995. **23**(3): p. ii-iv.
4. Janin, J., *Welcome to CAPRI: A Critical Assessment of PRedicted Interactions.* Proteins: Structure, Function, and Bioinformatics, 2002. **47**(3): p. 257.
5. Moitessier, N., et al., *Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go.* British Journal of Pharmacology, 2008. **153**(S1): p. S7-S26.
6. Kufareva, I., et al., *Status of GPCR modeling and docking as reflected by community wide GPCR DOCK 2010 assessment.* Structure, 2011. **19**(8): p. 1108-26.
7. Michino, M., et al., *Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008.* Nat Rev Drug Discov, 2009. **8**(6): p. 455-63.
8. Jaakola, V.-P., et al., *The 2.6 Angstrom Crystal Structure of a Human A2A Adenosine Receptor Bound to an Antagonist.* Science, 2008. **322**(5905): p. 1211-17.
9. Chien, E.Y.T., et al., *Structure of the Human Dopamine D3 Receptor in Complex with a D2/D3 Selective Antagonist.* Science, 2010. **330**(6007): p. 1091-5.
10. Wu, B., et al., *Structures of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists.* Science, 2010. **330**(6007): p. 1066-71.
11. Cherezov, V., et al., *High-Resolution Crystal Structure of an Engineered Human beta2-Adrenergic G Protein-Coupled Receptor.* Science, 2007. **318**(5854): p. 1258-65.
12. Warne, T., et al., *Structure of a beta1-adrenergic G-protein-coupled receptor.* Nature, 2008. **454**(7203): p. 486-91.
13. Gloriam, D.E., et al., *Definition of the G Protein-Coupled Receptor Transmembrane Bundle Binding Pocket and Calculation of Receptor Similarities for Drug Design.* J Med Chem, 2009. **52**(14): p. 4429-42.
14. Lundstrom, K., et al., *Mapping of Dopamine D3 Receptor Binding Site by Pharmacological Characterization of Mutants Expressed in Cho Cells with the Semliki Forest Virus System.* Journal of Receptors and Signal Transduction, 1998. **18**(2-3): p. 133-50.
15. Sartania, N. and P.G. Strange, *Role of Conserved Serine Residues in the Interaction of Agonists with D3 Dopamine Receptors.* Journal of Neurochemistry, 1999. **72**(6): p. 2621-4.
16. Shi, L. and J.A. Javitch, *The Binding Site of Aminergic G protein-Coupled Receptors: The Transmembrane Segments and Second Extracellular Loop.* Annual Review of Pharmacology and Toxicology, 2002. **42**(1): p. 437-67.
17. Thoma, G., et al., *Orally Bioavailable Isothioureas Block Function of the Chemokine Receptor CXCR4 In Vitro and In Vivo.* J Med Chem, 2008. **51**(24): p. 7915-20.
18. Rosenkilde, M.M., et al., *Molecular Mechanism of Action of Monocyclam Versus Bicyclam Non-peptide Antagonists in the CXCR4 Chemokine Receptor.* J Biol Chem, 2007. **282**(37): p. 27354-65.
19. Rosenkilde, M.M., et al., *Molecular Mechanism of AMD3100 Antagonism in the CXCR4 Receptor.* J Biol Chem, 2004. **279**(4): p. 3033-41.
20. Wong, R.S.Y., et al., *Comparison of the Potential Multiple Binding Modes of Bicyclam, Monocylam, and Noncyclam Small-Molecule CXC Chemokine Receptor 4 Inhibitors.* Mol Pharmacol, 2008. **74**(6): p. 1485-95.
21. DeMarco, S.J., et al., *Discovery of novel, highly potent and selective [beta]-hairpin mimetic CXCR4 inhibitors with excellent anti-HIV activity and pharmacokinetic profiles.* Bioorganic & Medicinal Chemistry, 2006. **14**(24): p. 8396-404.
22. Moncunill, G., et al., *Anti-HIV Activity and Resistance Profile of the CXC Chemokine Receptor 4 Antagonist POL3026.* Mol Pharmacol, 2008. **73**(4): p. 1264-73.
23. Ballesteros, J. and H. Weinstein, *Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations of G protein-coupled receptors.* Methods Neurosci, 1995. **25**: p. 366-428.
24. Surgand, J.S., et al., *A chemogenomic analysis of the transmembrane binding cavity of human G protein-coupled receptors.* Proteins: Structure, Function, and Bioinformatics, 2006. **62**(2): p. 509-38.

25. Vroling, B., et al., *GPCRDB: information system for G protein-coupled receptors.* Nucleic Acids Res, 2011. **39**(suppl 1): p. D309-D319.
26. Sanders, M.P., et al., *ss-TEA: Entropy based identification of receptor specific ligand binding residues from a multiple sequence alignment of class A GPCRs.* BMC Bioinformatics, 2011. **12**(1): p. 332.
27. Consortium, T.U., *The Universal Protein Resource (UniProt) in 2010.* Nucleic Acids Research, 2010. **38**(suppl 1): p. D142-D148.
28. Larkin, M.A., et al., *Clustal W and Clustal X version 2.0.* Bioinformatics, 2007. **23**(21): p. 2947-8.
29. Scheerer, P., et al., *Crystal structure of opsin in its G-protein-interacting conformation.* Nature, 2008. **455**(7212): p. 497-502.
30. Palczewski, K., et al., *Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor.* Science, 2000. **289**(5480): p. 739-45.
31. Xu, F., et al., *Structure of an Agonist-Bound Human A2A Adenosine Receptor.* Science, 2011. **332**(6027): p. 322-7.
32. Rasmussen, S.G.F., et al., *Structure of a nanobody-stabilized active state of the beta2 adrenoceptor.* Nature, 2011. **469**(7329): p. 175-80.
33. Rasmussen, S.G.F., et al., *Crystal structure of the beta2 adrenergic receptor-Gs protein complex.* Nature, 2011. **doi: 10.1038/nature10361**.
34. Deupi, X., et al., *Structural Models of Class A G Protein-Coupled Receptors as a Tool for Drug Design: Insights on Transmembrane Bundle Plasticity* Current Topics in Medicinal Chemistry, 2007. **7**(10): p. 991-8.
35. Deupi, X., et al., *Ser and Thr Residues Modulate the Conformation of Pro-Kinked Transmembrane α-Helices.* Biophysical Journal, 2004. **86**(1): p. 105-15.
36. de Graaf, C., et al., *Molecular modeling of the second extracellular loop of G-protein coupled receptors and its implication on structure-based virtual screening.* Proteins: Structure, Function, and Bioinformatics, 2008. **71**(2): p. 599-620.
37. Goldfeld, D.A., et al., *Successful prediction of the intra- and extracellular loops of four G protein-coupled receptors.* Proc Natl Acad Sci, 2011. **108**(20): p. 8275-80.
38. Mehler, E.L., et al., *Ab initio computational modeling of loops in G-protein-coupled receptors: Lessons from the crystal structure of rhodopsin.* Proteins: Structure, Function, and Bioinformatics, 2006. **64**(3): p. 673-90.
39. Nikiforovich, G.V., et al., *Modeling the possible conformations of the extracellular loops in G protein-coupled receptors.* Proteins: Structure, Function, and Bioinformatics, 2010. **78**(2): p. 271-85.
40. Moe (The Molecular Operating Environment) Version 2009.10 Chemical Computing Group Inc. Montreal Canada http:www//www.chemcomp.com.
41. Berkhout, T.A., et al., *CCR2: Characterization of the Antagonist Binding Site from a Combined Receptor Modeling/Mutagenesis Approach.* J Med Chem, 2003. **46**(19): p. 4070-86.
42. Jensen, P.C., et al., *Molecular Interaction of a Potent Nonpeptide Agonist with the Chemokine Receptor CCR8.* Mol Pharmacol, 2007. **72**(2): p. 327-40.
43. Rosenkilde, M.M., et al., *Activation of the CXCR3 Chemokine Receptor through Anchoring of a Small Molecule Chelator Ligand between TM-III, -IV, and -VI.* Mol Pharmacol, 2007. **71**(3): p. 930-41.
44. Govaerts, C.d., et al., *The TXP Motif in the Second Transmembrane Helix of CCR5.* J Biol Chem, 2001. **276**(16): p. 13217-25.
45. Case, D.A., et al., AMBER11, University of California, San Franscisco, USA, 2010
46. Lovell, S.C., et al., *Structure validation by Cα geometry: ϕ,ψ and Cβ deviation.* Proteins: Structure, Function, and Bioinformatics, 2003. **50**(3): p. 437-50.
47. Dunbrack, R.L. and F.E. Cohen, *Bayesian statistical analysis of protein side-chain rotamer preferences.* Protein Science, 1997. **6**(8): p. 1661-81.
48. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Research, 2000. **28**(1): p. 235-42.
49. Krieger, E., et al., *Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8.* Proteins: Structure, Function, and Bioinformatics, 2009. **77**(S9): p. 114-22.
50. Canutescu, A.A., A.A. Shelenkov, and R.L. Dunbrack, *A graph-theory algorithm for rapid protein side-chain prediction.* Protein Science, 2003. **12**(9): p. 2001-14.
51. Hooft, R.W.W., et al., *Errors in protein structures.* Nature, 1996. **381**(6580): p. 272-2.

52. Eswar, N., et al., *Comparative Protein Structure Modeling Using Modeller*, in *Current Protocols in Bioinformatics*. 2002, John Wiley & Sons, Inc.
53. Veldkamp, C.T., et al., *Structural Basis of CXCR4 Sulfotyrosine Recognition by the Chemokine SDF-1/CXCL12.* Science Signaling, 2008. **1**(37): p. ra4.
54. Kristiansen, K., *Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G-protein-coupled receptors: molecular modeling and mutagenesis approaches to receptor structure and function.* Pharmacology & Therapeutics, 2004. **103**(1): p. 21-80.
55. Ye, S., et al., *Tracking G-protein-coupled receptor activation using genetically encoded infrared probes.* Nature, 2010. **464**(7293): p. 1386-9.
56. Bokoch, M.P., et al., *Ligand-specific regulation of the extracellular surface of a G protein-coupled receptor.* Nature, 2010. **463**(7277): p. 108-12.
57. Huber, T. and T.P. Sakmar, *Escaping the flatlands: new approaches for studying the dynamic assembly and activation of GPCR signaling complexes.* Trends in Pharmacological Sciences, 2011. **32**(7): p. 410-19.
58. Ballesteros, J.A., L. Shi, and J.A. Javitch, *Structural Mimicry in G Protein-Coupled Receptors: Implications of the High-Resolution Structure of Rhodopsin for Structure-Function Analysis of Rhodopsin-Like Receptors.* Mol Pharmacol, 2001. **60**(1): p. 1-19.
59. Torrice, M.M., et al., *Probing the role of the cation-pi interaction in the binding sites of GPCRs using unnatural amino acids.* Proc Natl Acad Sci, 2009. **106**(29): p. 11919-24.
60. Varady, J., et al., *Molecular Modeling of the Three-Dimensional Structure of Dopamine 3 (D3) Subtype Receptor: Discovery of Novel and Potent D3 Ligands through a Hybrid Pharmacophore- and Structure-Based Database Searching Approach.* J Med Chem, 2003. **46**(21): p. 4377-92.
61. Sanders, M.P.A., et al., *Snooker: A structure-based pharmacophore generation tool applied to class A GPCRs*, 2011, **51**(9), p. 2277-92.
62. Warr, W., *ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI).* J Comput Aided Mol Des, 2009. **23**(4): p. 195-8.
63. Liu, X., et al., *Cyndi: a multi-objective evolution algorithm based method for bioactive molecular conformational generation.* BMC Bioinformatics, 2009. **10**(1): p. 101.
64. Nabuurs, S.B., M. Wagener, and J. de Vlieg, *A Flexible Approach to Induced Fit Docking.* J Med Chem, 2007. **50**(26): p. 6507-18.
65. Verdonk, M.L., et al., *Improved protein–ligand docking using GOLD.* Proteins: Structure, Function, and Bioinformatics, 2003. **52**(4): p. 609-23.
66. Angel, T.E., M.R. Chance, and K. Palczewski, *Conserved waters mediate structural and functional activation of family A (rhodopsin-like) G protein-coupled receptors.* Proc Natl Acad Sci, 2009. **106**(21): p. 8555-60.
67. Pardo, L., et al., *The Role of Internal Water Molecules in the Structure and Function of the Rhodopsin Family of G Protein-Coupled Receptors.* ChemBioChem, 2007. **8**(1): p. 19-24.
68. Katritch, V., et al., *Structure-Based Discovery of Novel Chemotypes for Adenosine A2A Receptor Antagonists.* J Med Chem, 2010. **53**(4): p. 1799-809.
69. Roberts, B.C. and R.L. Mancera, *Ligand-Protein Docking with Water Molecules.* J Chem Inf Model, 2008. **48**(2): p. 397-408.
70. Verdonk, M.L., et al., *Modeling Water Molecules in Protein-Ligand Docking Using GOLD.* J Med Chem, 2005. **48**(20): p. 6504-15.
71. Shimamura, T., et al., *Structure of the human histamine H1 receptor complex with doxepin.* Nature, 2011. **475**(7354): p. 65-70.

CHAPTER 7

A prospective cross-screening study on G protein-coupled receptors: lessons learned in virtual compound library design

Marijn P.A. Sanders[1], Luc Roumen[2], Eelke van der Horst[3], J. Robert Lane[3], Henry F. Vischer[2], Jody van Offenbeek[2,4], Henk de Vries[3], Stefan Verhoeven[4], Ken Y. Chow[2], Folkert Verkaar[2,4], Margot W. Beukers[3], Ross McGuire[4], Rob Leurs[2], Ad P. IJzerman[3], Jacob de Vlieg[1,4], Iwan J.P. de Esch[3], Guido J. Zaman[4], Jan P.G. Klomp[4], Chris de Graaf[2] and Andreas Bender[3]

[1]CMBI, NCMLS, Radboud University Nijmegen Medical Centre, Nijmegen, the Netherlands; [2]Department of Medicinal Chemistry, VU University Amsterdam, Amsterdam, the Netherlands; [3]LACDR, Division of Medicinal Chemistry, Leiden University, Leiden, The Netherlands; [4]Department of Molecular Design & Informatics, MSD, Oss, The Netherlands [5]Department of Molecular Pharamcology & DMPK, MSD, Oss, The Netherlands

in preparation

Acknowledgements

This work was supported by the Dutch Top Institute Pharma, project number: D1-105 and Netherlands Organization for Scientific Research (NWO) through a VENI grant (700.59.408 to C. de G.).

Abstract

For the first time, and prospectively, we investigated the performance of a receptor-based and a ligand-based virtual screening approach against a set of three G Protein-Coupled Receptors (GPCRs), namely the Beta-2 Adrenoreceptor (ADRB2), the Adenosine A2 Receptor (AA2AR) and the Sphingosine 1-Phosphate Receptor (S1PR1). This study aims to evaluate ligand- and structure-based approaches to make most use of information available today. In order to also evaluate cross-reactivity of ligands, all 900 compounds selected for any of the receptors (300 each) were also screened against the two other receptors that formed part of this study. Novel bioactive compounds could be identified using a consensus scoring procedure combining ligand-based and structure-based tools. The success of retrospective ligand-based and prospective ligand-protein based screening appeared to be dependent on the chemical diversity of the ligand training set used to validate and optimize the ligand- and protein-based models. While the diverse AA2AR ligand training set facilitated the construction of a robust ligand-based model superior to the protein-based model, a combination of ligand- and protein-based ADRB2 models gave the best results in retrospective screening runs. Furthermore, our in vitro screening studies suggest to be careful with restrictive in silico compound similarity filters to identify novel ligands. Finally, we found that there was a striking degree of activity of ligands that were selected for one receptor, and identified to be active on another receptor of the set. This effect could be partly explained by the fuzziness and overlap of protein-based pharmacophore models. Overall, this is one of the first prospective chemogenomics studies available in the literature in which all in silico hits are measured against all protein targets . The lessons learned from this exercise can be used to guide future virtual ligand library design efforts.

7.1 Introduction

Chemogenomics is a new research area aimed at systematically studying the biological effect of a wide variety of small molecules (ligands) on a wide variety of macromolecular targets (gene products) [1-4]. Experimental measurements of large quantities of ligands and targets are time consuming and cost intensive and are therefore often complemented by high-throughput in silico chemogenomics approaches. These methods are typically divided in ligand- and target-based approaches and profile either a ligand against a set of diverse proteins, or a set of ligands against one specific protein target [5, 6] Both methods have been successfully applied in virtual screening experiments [5, 6] to guide rational drug discovery and design [7-10]. However, the applicability domain of these in silico chemogenomics approaches depends on the quality and completeness of the training sets used for model construction and validation [11]. Data on small molecules is often incomplete since molecules are usually not screened systematically through a large panel of protein targets but only on a few pharmacological interesting targets [12]. Furthermore, most scientific studies focus on the presentation of "active" molecules and usually are (potentially) "inactive" molecules not synthesized/tested (e.g., in hit optimization studies) or not reported [12]. Even in target annotated ligand databases such as CheMBLdb [13], DrugBank [14], BindingDB [15-17], PDBBind [18, 19], MOAD [20, 21], WOMBAT [22] and Glida-DB [23], protein-ligand interaction matrices are incomplete [12]. Computational methods have however been successfully used to fill the gap in experimental ligand-target affinity matrices[24], and to identify new drug-target associations [5, 25]. Furthermore, the high number of active molecules in target annotated chemical databases allows the identification of molecular features that determine binding to specific proteins and protein classes [26].

Ligand-based virtual screening methods like substructure mining [26], molecular fingerprint similarity searches [29] and ligand-based pharmacophore models [30] are generally faster than target(-ligand interaction) based methods such as molecular docking [31] and protein structure-based pharmacophore models [32]. Structure-based methods on the other hand are more suitable to find novel ligands and offer insight in the atomic details of protein-ligand interactions in 3D. The latter methods are therefore more suitable for the design of novel and diverse sets of bioactive molecules, but are generally too slow for efficient navigation of chemogenomics space. The recent elucidation of GPCR structures enables in silico screens based on structure-based approaches for this protein family [33], including protein-based pharmacophore screening methods that use the spatial arrangement of key chemical features required for protein-ligand binding, to identify new ligands and predict their binding mode in the protein target [34].

The research project described in this paper is an effort to merge the current chemogenomics thinking of multi-target bioactivities with the increased availability of X-ray structures of GPCRs as well as the increased availability of chemogenomics databases such as ChEMBL. Accordingly, we applied two different virtual screening

approaches, one based on the recently elucidated receptor structures and one based on ligand bioactivity information, in order to select potentially bioactive compounds against a panel of three receptors, namely the Beta-2 Adrenoreceptor (ADRB2), the Adenosine A2 Receptor (AA2AR) and the Sphingosine 1-Phosphate Receptor (S1PR1). While ADRB2 plays an important role in cardiovascular disorders [35] and asthma [35] , AA2AR is involved in coronary disease [35] and parkinson's disease [36], and S1PR1 is an important target in the treatment of autoimmune diseases [35] and potentially cancer [37]. ADRB2 and AA2AR were the first druggable GPCRs for which crystal structures were solved [38, 39], while the first crystal structure of S1PR1 was announced recently [40]. Hence, we selected this receptor set due to a combination of biological relevance, the availability of crystal structure information (X-ray structures for ADRB2 and AA2AR, not yet for S1PR1) as well as the availability of known bioactive ligands (large, chemically diverse (AA2AR), large, chemically similar (ADRB2), and small, chemically similar (S1PR1) ligand sets), and also due to our ability to perform experimental assays for those targets to validate our models.

We selected a large number of compounds (300 per receptor) in a prospective manner and tested them for bioactivity against their intended target, but also against the other receptors in the study. This approach allowed us to compare structure-based and ligand-based screening approaches in the context of currently available structural and bioactivity information; the degree of cross-reactivity of ligands against the set of receptors; and to explore whether information from related receptors might be useful in a practical, prospective virtual screening setting.

7.2 Methods

7.2.1. Pharmacophore Screening

Figure 7.1 depicts a schematic overview of the applied virtual screening protocol. Structure based pharmacophores describing the negative image of the pocket, situated in between the transmembrane helices, were generated for all three GPCRs with Snooker [41]. For AA2AR and ADRB2 a consensus model of eight different templates was constructed whereas for S1PR1 a model was generated based on an experimentally validated agonist-bound homology model [42]. To gain specificity, directionality was added to the polar pharmacophore features as the average vector between the polar feature centre and the Cα-atom of the residues which constitute the respective features.

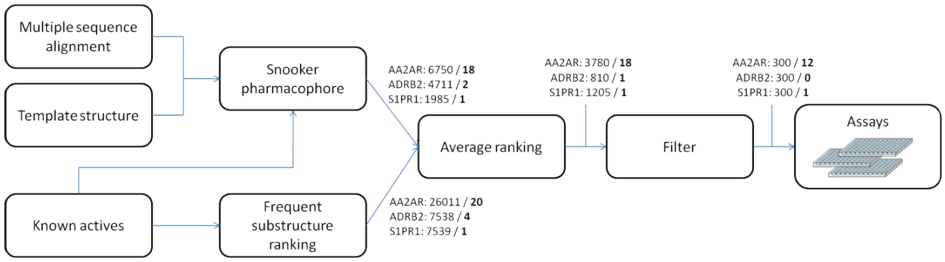


Figure 7.1: Virtual screening flowchart. Numbers indicate the number of compounds which pass the respective structure-based and ligand-based bioactivity models, bold numbers indicate the number of actives included in the selections. Note that out of the 300 compounds selected for each of the targets, some overlap between the hit lists had to be removed and not all compounds were available as physical samples, leading to slight modifications of the hit lists suggested by the virtual screening methods.

7.2.2. Ligand training set used for pharmacophore generation

Known active compounds for the three target proteins investigated in this study, namely the Beta-2 Adrenoreceptor (ADRB2), the Adenosine A2 Receptor (AA2AR) and the Sphingosine 1-Phosphate Receptor (S1PR1), were extracted for all species from the ChEMBL database (Release August 2009)[13] independently of the functional class (agonist, partial agonist, antagonist, inverse agonist) using an activity cutoff on Ki, IC50, or EC50 of less than 50nM. To reduce the bias introduced by the deposition of compound series, diverse subsets of 50 compounds were generated for the AA2AR and ADRB2 receptor by exclusion sphere clustering on Tanimoto distances between the BCI fingerprints of the compounds. For the S1PR1 receptor all 43 compounds reported in ChEMBL were used as training set.

7.2.3. Pharmacophore model generation and training

Structure-based pharmacophores for all three targets were subsequently trained with their corresponding sets of known actives. In this training, the subsets of pharmacophore features were ranked first according to the number of features in the pharmacophore and next according to the number of known actives that they were able retrieve. Finally, shape restraints were defined by the Tanimoto distance of the compounds matching the ADRB2 and AA2AR pharmacophores to carazolol and ZMA, respectively, which are co-crystallized in their respective receptors [38, 39]. For S1PR1, shape restraints included the Tanimoto distance of compounds matching the pharmacophore to the average pose of training set actives in this pharmacophore. Last calculated as the pose with the lowest average Tanimoto distance to all poses matching the pharmacophore. Tanimoto distance cutoffs were set at those values at which the enrichment of known actives over the 50308 database compounds was optimal.

7.2.4. Pharmacophore screening settings

Pharmacophore screening was performed with RDKit using the procedure described by Sanders et al. [41]. Projected points were also calculated with RDKit and handled just as the other feature types, with the exception that a heavy-atom – projected-point pair cannot be separated and has to match a feature projected-feature pair. The angular difference between the vectors describing the heavy-atom projected point pair and the feature projected feature pair has to be within 45 degrees.

7.2.5. Preparation prospective pharmacophore screening library

The compound library used for screening (**paragraph 7.2.10**) was prepared as follows. Initial three-dimensional conformations were generated with Corina [43] and multiple 3D conformations were created with a genetic algorithm, Cyndi [44], which was ran with a population size of 200 and final output of 100 conformations per molecule. Given that the algorithm used here is designed for screening millions of compounds of typical in-house or vendor libraries, no further force-field minimization of compounds has been performed.

7.2.6. Frequent substructure ranking

For the ligand-based screening part, the substructure-based screening method of van der Horst et al. [45] has been employed in the current study. This method performs screening by searching library compounds for the occurrence of specific substructures. These substructures are derived from existing ligands of the target under study and are selected for their ability to distinguish ligands for this target from other molecules.

To assemble sets of ligands (the source sets) for the human adenosine A2A receptor (AA2AR), β¬2 adrenoceptor (ADRB2), and the S1P¬1¬ lysophospholipid receptor (S1PR1), ligand structures and activity data were retrieved from the ChEMBL database [46], selecting compounds with activity, i.e. a Ki, IC50, or EC50, of 10nM or less for the adenosine A2A receptor, and 10µM or less for the other two receptors. The 43 reference ligands used for fine-tuning the structural models of the S1PR1 receptor were also included in the training sets. Subsequent manual inspection was performed to ensure further removal of any alleged agonists, for instance, compounds that were highly similar to adenosine. All source sets were split into a training set and a test set using Pipeline Pilot's Diverse Molecules component (30% test set and FCFP_4 fingerprints). The set sizes are provided in **Table 7.1**. For analysis of the substructures, all training sets were contrasted against a background set of 'average' compounds (the background set). The background set consisted of 10,000 randomly selected compounds from the drug-like subset of the ZINC database (accessed: February 12, 2010), a collection of all available chemical compounds [47]. Chemical structures were represented as graphs using a special type for aromatic bonds. Elaborate chemical representations, described in detail in the work of Kazius et al. [48] and van der Horst et al. [26] were not used.

Table 7.1: Number of molecules in training and test set for each receptor for the substructure-based compound ranking method.

	Training set	Test set	Total
AA2AR	179	76	255
ADRB2	301	129	430
S1PR1	172	74	246

7.2.7. Generation of Frequently Occurring Substructures

Frequent substructure sets were generated using the frequent graph miner Gaston which finds all possible frequently occurring substructures in a set of molecules [49]. For each substructure, the number of molecules the substructure occurred in was calculated. The difference between the relative occurrence (fraction) of a substructure in the antagonists set and the background set is the score contribution of that substructure. Substructures were ranked according to the score contribution in descending order. The top 50 best substructures were selected for the screening model.

7.2.8. Substructure-based Virtual Screening – Ranking of Compounds.

To rank compounds in order of likelihood to display receptor binding, a score was calculated based on the previously generated substructure set. The score for a compound was calculated as follows: for each substructure in the set, presence in the compound was determined. For the substructures that occurred in a compound, the score contribution was summed to calculate the final score for that compound.

7.2.9. Small Scale Benchmark Screening.

All screening models were benchmarked using the test sets that were reserved earlier. Receiver operator curves (ROC) were plotted and the area under the curve (AUC) was calculated with Pipeline Pilot 6.1.5.0 Student Edition. The substructure set and score calculation that resulted in the highest AUC, was selected for the large-scale virtual screening.

7.2.10. Virtual Screening Library

A diverse subset of the MSD in-house library of 50,308 compounds was used for virtual screening. In order to characterize the library, the overlap of the MSD in-house library with the ZINC purchasable compound set (~23,7M) was determined. The latter consisted of compounds from 26 commercial vendors, with a MW ≤ 500 Dalton. To determine the overlap between the two sets, structures were converted into unique hash codes without considering stereochemistry. From the 50,308 MSD compounds, 85% occurred in the 23,691,219 ZINC [47] compound database and 60% occurred in the 6,981,556 CoCoCo [50] compound database. The MSD compounds possess similar physico-chemical properties to those within the ZINC and CoCoCo compound databases, placing emphasis

on druglikeness [51, 52] (data not shown). Compounds that occurred in the training or test sets were removed from the screening library, as well as compounds that had already been tested against one of the targets (for human, mouse, and rat). In addition, for the AA2A receptor, compounds with a typical AR ligand scaffold, such as xanthines, were removed.

7.2.11. Compound selection

To focus on new chemical entities we removed for each selection the compounds with a similarity in ECFP_4 / Tanimoto space > 0.5 to a known active as well as compounds with known AA2AR scaffolds in the AA2AR selection. A selection of 300 compounds for each receptor was made after determination of the average rank of each compound which matched a pharmacophore and had a positive score in the frequent substructure procedure. In case that compounds were not available for testing, the next best compound was selected.

7.2.12. Experimental validation

7.2.12.1. Adenosine A2A Receptor

HEK293 cells stably expressing the human AA2AR receptor (gift from Dr. Wang, Biogen, Cambridge, MA) were used to determine the affinity of compounds in a radioligand binding assay with [^{3}H]ZM241385 as the radioligand. Membranes containing 40 µg protein were incubated in a total volume of 100 µL Tris•HCl (50 mM, pH 7.4) and [^{3}H]ZM241385 (final concentration 1.7 nM) for 2 h at 25 ^{o}C in a shaking water bath. Nonspecific binding was determined in the presence of 100 µM CGS21680. The incubation was terminated by filtration over pre-wetted Whatman GF/B filters under reduced pressure with a Brandel harvester. Filters were washed three times with ice-cold buffer and placed in scintillation vials. Emulsifier Safe (3.5 mL) was added, and after 2 h radioactivity was counted in a TriCarb 2900TR liquid scintillation counter. Compounds that inhibited binding by ≥ 50% at 10 µM were subject to testing in concentration-response curves.

7.2.12.2. Beta-2 Adrenoreceptor Assay

HEK293T cells were cultured and transiently transfected with 2.5 µg ADRB2-pcDNA3.1+ (obtained from the Missouri S&T cDNA Resource Center) per 106 cells using 12 µg linear 25-kDa polyethylenimine (Polysciences, Warrington, PA, USA) as described previously[53]. Cells were harvested 48h after transfection and membrane fractions were prepared as described previously[53]. ADRB2-expressing membranes were incubated at room temperature in 96-well plates in binding buffer (50 mM HEPES – pH 7.4, 1 mM CaCl2, 5 mM MgCl2, 100 mM NaCl, and 0.5% (w/v) BSA) with 1 nM [3H]-dihydroalprenolol (DHA; 104.4 Ci/mmol from PerkinElmer Life Sciences), and 10 µM or increasing concentrations of compounds. After 1h, incubations were terminated by rapid filtration through Unifilter GF/C plates (PerkinElmer Life Sciences) presoaked in 0.5%

polyethylenimine and washed with ice-cold binding buffer supplemented with 500 mM NaCl. Radioactivity was measured using a MicroBeta Trilux (PerkinElmer Life Sciences). Nonlinear regression analysis of data and calculation of Kd and Ki values was performed using GraphPad Prism 4 software.

7.2.12.3. Sphingosine-1-Phosphate Receptor

The assay was performed using the Pathhunter™ Enzyme Fragment Complementation β-arrestin recruitment technology as described previously for the S1PR1 receptor.[54] The PathHunter™ CHO-K1 EDG1 β-arrestin EFC cell line was purchased from DiscoveRx (Fremont, CA). Cells were cultured in Dulbecco's modified Eagles medium F-12 (Invitrogen, Garlsbad, CA), supplemented with 10% heat-inactivated fetal calf serum (Cambrex, Verviers, Belgium) , 100 U/mL penicillin, 100 µg/mL streptomycin, 300 µg/mL hygromycin B and 800 µg/mL geneticin (Invitrogen). Cells were seeded at a density of 10,000 cells per well of a 384 wells culture plates (PerkinElmer, Boston, Massachusetts) in 20 µL OPTI-MEM (Invitrogen). After overnight incubation at 37ºC in a humidified incubator (5% CO2, 95% humidity), 4 µL of compound dilution was added to cells and the plate was returned to the incubator for 2 hours, followed by incubation at room temperature for 1 hour. Cells were lysed using 8 µL PathHunter™ detection reagent (DiscoveRx). Plates were incubated in the dark for 2 hour at room temperature before measurement of β-galactosidase activity (chemiluminescence) on an Envision multilabel plate reader (PerkinElmer Life Sciences). Compounds that induced ≥ 30% β-arrestin recruitment compared to the reference compound AUY954 were selected and tested in dose response curves.

7.3. Results and discussion

This aim of the current study was to evaluate the performance of (combined) ligand- and structure-based virtual screening approaches (**Figure 7.1**), optimized using retrospective screening simulations (**Figure 7.2** and **7.3**), and applied in prospective all-against-all GPCR chemogenomics studies (**Figure 7.4** and **7.5**, Tables **7.1** and **7.2**).

7.3.1. Chemical diversity ligand training sets

A large diversity space of active compounds is desired in drug design projects because it allows the selection and synthesis of drug-like compounds with good solubility, ADMET properties, selectivity towards the targets and a strong intellectual property position. The diversity of the 50 most dissimilar compounds for AA2AR and ADRB2 in BCI fingerprint space is presented in **Figure 7.2**. ADRB2 compounds show to be more similar to each other (average Euclidean distance ~11) and form a tight cluster as compared to AA2AR compounds (average Euclidean distance ~14). This indicates that the chance of finding novel compounds for the AA2AR receptor is larger than for the ADRB2 receptor, especially if we take into account that all compounds that are similar to known actives have been removed.

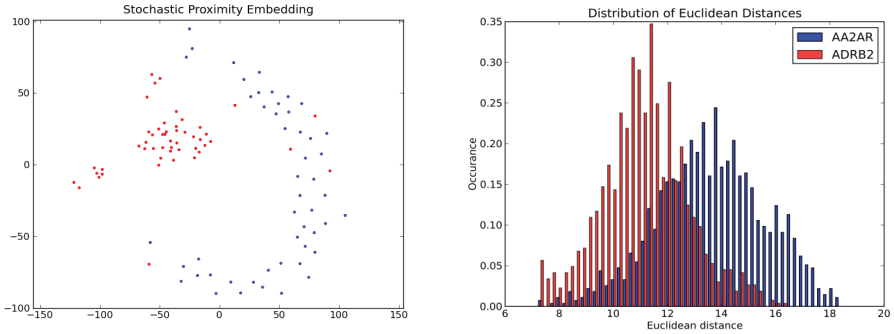


Figure 7.2: Visualization of ligand similarities for the known actives on the adenosine A2A and the beta-2 adrenergic receptor. While, on average, beta-adrenoceptor ligands are more similar to each other, also some atypical ligands can be found which resemble compounds active on the adenosine A2A receptor. We will revisit this similarity of bioactivity classes later in the text, when we discuss the surprising degree of cross-reactivity of ligands selected for one receptor, but found to be bioactive also (or solely) against other receptors of the set.

7.3.2. Pharmacophore models match experimentally supported ligand binding modes

Several polar residues, namely T3.36, Q3.37, N6.55, E45.53 and S7.42 make H-bond interactions with co-crystallized AA2AR ligands in crystal structures[39, 55] , and are shown to be important in agonists and xanthine antagonist binding based on site-directed mutagenesis studies.[56, 57], [58], [59], [39, 60] The structure-based pharmacophore model of AA2AR used in our virtual screening studies includes interaction features derived from three of these polar residues, namely T3.36, N6.55 and S7.42 (Figure 7). ADRB2 ligands share an essential positively charged amine as well as an aromatic ring separated by circa 5Å (partial and full agonists) to 7Å (inverse agonists and antagonists). The ADRB2 crystal structure [38] shows protein-ligand interactions to residues D3.32, S5.42, N7.39 and F6.52, which is in line with site-directed mutagenesis studies.[61, 62], [63], [64] Furthermore, mutation of residues S5.43, S5.46, N6.55, and Y7.35 have been shown to affect partial and full agonist binding[65-67], 35. The ADRB2 pharmacophore model indeed contains a hydrophobic contact with V3.33 and polar interactions with N7.39, D3.32, S5.42 and S5.46 as supported by ligand co-crystallized ADRB2 crystal structures and site-directed mutagenesis studies (**Figure 7.6 and 7.7**). S1PR1 receptor ligands are characterized by the presence of a polar head which contains negatively and positively charged groups and a long hydrophobic tail[68]. Based on these chemical ligand properties computational modeling and site-directed mutagenesis studies have identified important S1PR1 receptor-ligand interactions including ionic interactions between the negatively charged phosphate oxygens of S1P and R3.28, and between the protonated amine of S1P and the E3.29.[69, 70] In addition, Y5.39 has been identified as a conserved feature to influence selectivity for the lysophosphatidic acid receptor subtypes [71]. The structure-based pharmacophore model for S1PR1 indeed includes features for R3.28, E3.29 and Y5.39 as supported by site-directed mutagenesis studies (**Figure 7.6 and 7.7**).

7.3.3. Retrospective virtual screening validation

A retrospective virtual screen based on the ligand-based, structure-based and the combined compound selection method of the 50 diverse actives for AA2AR and ADRB2 versus the 50308 assumed inactive compounds of the used compound library showed significant early enrichment (**Figure 7.3**). Retrospective enrichment for AA2AR is poor for the structure-based method. Probably due to the fact that a large portion of AA2AR ligands bind in the extracellular domain that is not represented in Snooker pharmacophores, and because it is likely that the A2A adenosine receptor has no family conserved receptor binding pocket[39]. The ligand-based method on the other hand has an excellent performance for this target receptor, but might be biased towards known chemistry. A combination of both methods is likely to result in a reduced number of identified actives as compared to only the substructure method, but might contain the desired novelty amongst identified actives, containing novel scaffolds and potentially a different set of interactions with the receptor than training set molecules. In contrast to AA2AR ligands, the ADRB2 ligands bind largely within the TM domain and share a common binding mode within a buried pocket. Therefore, structure based searches perform better on ADRB2 than on AA2AR. They are not only able to capture the actives faster, but also retrieve a higher percentage of actives. A combination of structure-based and ligand-based methods for this receptor outperforms both individual methods (**Figure 7.3**).

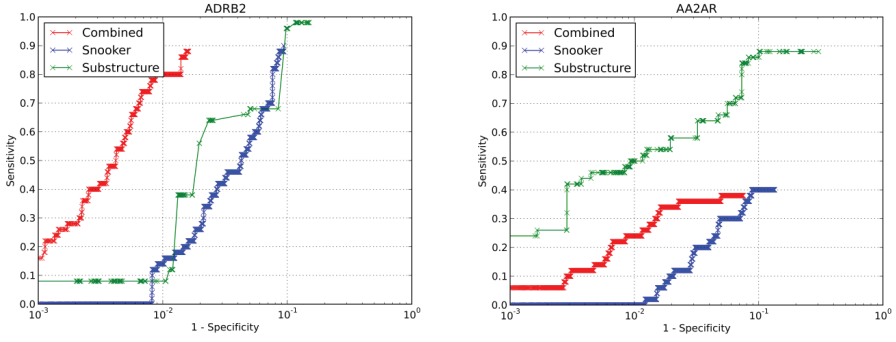


Figure 7.3: Receiver Operating (ROC) curves for the ligand-based and structures based virtual screening methods employed in this work. While the substructure-based method, on average, is able to achieve higher enrichment than the structure-based method employed, the extent of this difference is very much dependent on the target considered, with the adenosine A2A receptor showing significantly higher retrieval of active compounds than the beta-2 adrenoceptor. Consensus scoring outperforms each individual method in case of the beta-2 adrenoceptor.

7.3.4. Prospective cross-screening

Using each of the three bioactivity models a total of 300 compounds were selected. Duplicates were removed and in case the selected compound was not available anymore, the next compound in the list was selected. The final selection of compounds is

visualized in **Figure 7.4** and shows that the ADRB2 model and the S1PR1 models select the same compounds more often than other pairs of predictive models. Hence, overlap between all three bioactivity classes could be anticipated also in compound selection space; however, correlation between overlap in 'selection space' and experimentally confirmed cross-reactivity between receptors was not entirely correlated as described in more detail below.

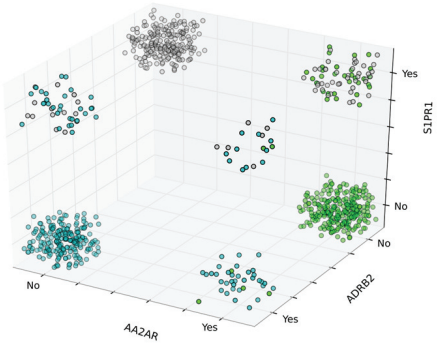


Figure 7.4: Distribution of compounds selected for experimental screening by each of the in silico bioactivity models. It can be seen that the beta-2 adrenoreceptor model and the sphingosine-1-phosphate receptor model select the same compounds more often than other pairs of activity models given the fact that 3780, 810 and 1205 compounds are scored for the AA2AR, ADRB2 and S1PR1 receptor model, respectively.

Using our combined ligand- and structure-based virtual screening approach, we have successfully identified 18 AA2AR, 6 ADRB2 and 3 S1PR1 ligands (**Figure 7.5**). Structures of the new compounds are visualized in Table 2 and plots of compound ranks versus receptor activity are displayed in **Figure 7.5**, also distinguishing between the in silico models used to select each bioactive compound identified in this study.

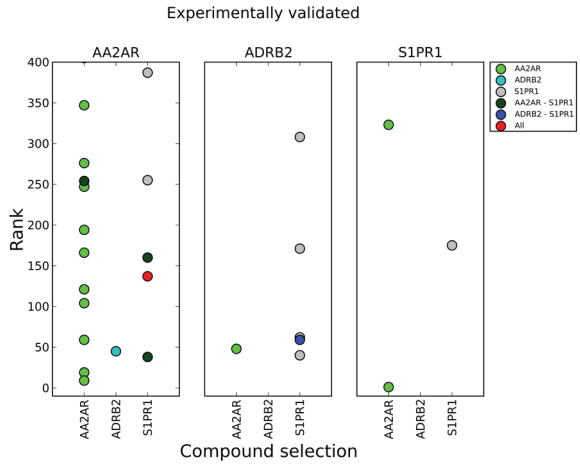
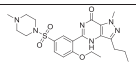
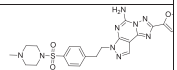
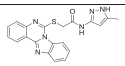
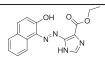
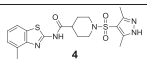
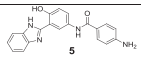
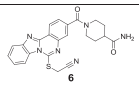
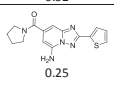
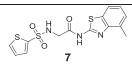
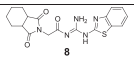
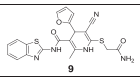
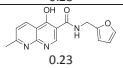
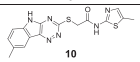
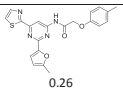
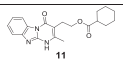
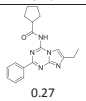
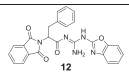
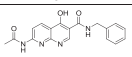
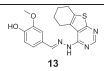
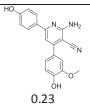
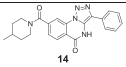
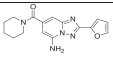
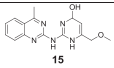
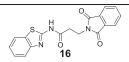
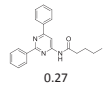
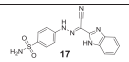
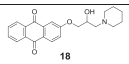
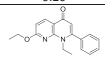
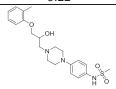
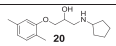
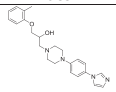
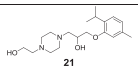
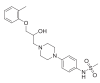
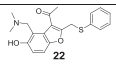
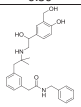
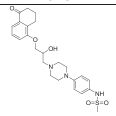
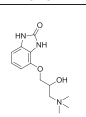
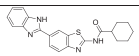
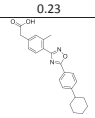
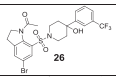
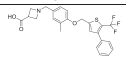
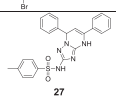
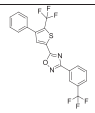


Figure 7.5: Novel active compounds found using the respective models for the three receptors in this study (adenosine A2 receptor, beta-2 adrenergic receptor and sphingosine-1-phosphate receptor). The title of the plot shows the receptor for which the ligands were experimentally found to be active against, while the bottom of the plot shows the model that was used in silico to select each respective compound.

Table 7.2: Identified active compounds for the AA2AR, ADRB2 and S1PR1 receptor. The last column of the table lists the closest available compound reported in the ChEMBL database (as calculated using the Tanimoto distance in combination with ECFP_4 fingerprints).

Compound	Molecular Weight (Da)	Rank AA2AR	Rank ADRB2	Rank S1PR1	Activity on receptor	Binding affinity (pKi) / potency (pEC50)[a]	Most similar compound in ChEMBL
 1	475	2285	-	160	AA2AR	6.7	 0.26
 2	388	347	-	-	AA2AR	6.2	 0.29
 3	310	247	-	-	AA2AR	5.9	 0.26
 4	433	254	-	1110	AA2AR	6.0	 0.26
 5	344	405	-	-	AA2AR	5.8	 0.32
 6	445	-	-	255	AA2AR	6.0	 0.25
 7	367	121	-	-	AA2AR	5.8	 0.27
 8	385	59	-	-	AA2AR	5.3	 0.28
 9	452	19	-	-	AA2AR	5.2	 0.23
 10	370	9	-	-	AA2AR	5.4	 0.26
 11	353	-	-	387	AA2AR	5.6	 0.27
 12	453	657	415	137	AA2AR	5.4	 0.24
 13	354	166	-	-	AA2AR	5.3	 0.23
 14	387	3636	-	38	AA2AR	5.6	 0.28
 15	299	194	-	-	AA2AR	6.2	 0.28

Compound	Molecular Weight (Da)	Rank AA2AR	Rank ADRB2	Rank S1PR1	Activity on receptor	Binding affinity (pKi) / potency (pEC50)[a]	Most similar compound in ChEMBL
 16	351	104	-	-	AA2AR	5.3	 0.27
 17	340	276	-	-	AA2AR	5.3	 0.26
 18	365	-	45	-	AA2AR	5.3	 0.22
 19	291	-	-	40	ADRB2	6.1	 0.38
 20	263	-	-	171	ADRB2	5.7	 0.33
 21	336	-	624	59	ADRB2	4.4	 0.36
 22	355	-	-	62	ADRB2	4.4	 0.20
 23	353	-	-	308	ADRB2	3.1	 0.26
 24	205	48	-	-	ADRB2	4.6	 0.23
 25	376	1	-	-	S1PR1	4.7	 0.22
 26	547	-	-	175	S1PR1	4.5	 0.18
 27	444	323	-	-	S1PR1	4.3	 0.16

[a] For AA2AR and ADRB2 pKi values are reported obtained from competition binding assays measuring radio ligand displacement. For S1PR1 pEC50 values are reported obtained from functional assays by measurement of the chemiluminescense due to β-galactosidase activity.

The AA2AR selection method was very successful and selected 12 out of the 18 active compounds for this receptor (including compound 3, for which the proposed binding mode in AA2AR is depicted in **Figure 7.6**). Another 3 compounds were ranked for this receptor but tested because they had a better rank in another receptor. The existence of GPCRlike compounds[72] (less flexible, less polar and more hydrophobic) and privileged scaffolds is well-known and is probably also reflected in our selection procedure since actives were retrieved for AA2AR in as well the ADRB2 and S1PR1 compound selection.
For the ADRB2 receptor only 6 compounds were identified (including compound 21, for which the proposed binding mode in ADRB2 is shown in **Figure 7.6**), a much smaller number than for AA2AR. This is in agreement with the earlier observed restricted topological structure space of compounds for this receptor compared to the AA2AR receptor. The active compounds that were identified as ADRB2 bioactives in this study resulted from the compound selection for the AA2AR and S1PRR targets. The fact that we did not find any ADRB2 ligands from the ADRB2 test set might relate to the fact that all compounds similar to known ADRB2 actives were excluded from the compound set. Remarkably, 5 of the ADRB2 bioactives were retrieved with the S1PR1 receptor pharmacophore. Important hydrogen bonding and charged interactions are in both SBPs related to interactions in TM3 and TM5 and hydrophobic interactions originate in ADRB2 mainly from Val3.33 and in S1PR1 from 5.39 and 6.55 but are in both cases located at fairly similar positions. The match of compound 21 in the S1PR1 pharmacophore is shown in **Figure 7.7** including a fit of the crystal structure conformation of carazolol of the ADRB2 receptor[38] in this pharmacophore. The complementarities of the S1PR1 pharmacophore to known ADRB2 binders and high similarity of the S1PR1 pharmacophore to the ADRB2 pharmacophore explains the ability of the S1PR1 SBP to retrieve ADRB2 actives.
For S1PR1, three agonists were identified of which one was in the S1PR1 compound selection (including compound 26, for which the proposed binding mode in the S1PR1 binding pocket is presented in **Figure 7.6**). Like the known bioactives, the novel hits are elongated compounds with physico-chemical properties similar to the endogenous ligand. Overall, two of the three compounds active against S1PR1 were actually selected by the AA2AR model, indicating a degree of cross-reactivity in the current study.
The chemical similarity of the bioactives identified in our study to known inhibitors is portrayed by the ECFP4 closest similarity calculation in **Table 7.2**. The ECFP4 similarity thresholds for chemical novelty lies at values of 0.26 (strict) [73] and 0.4 (loose) [74]. From this we can surmise that many of our hits can indeed be considered as novel.

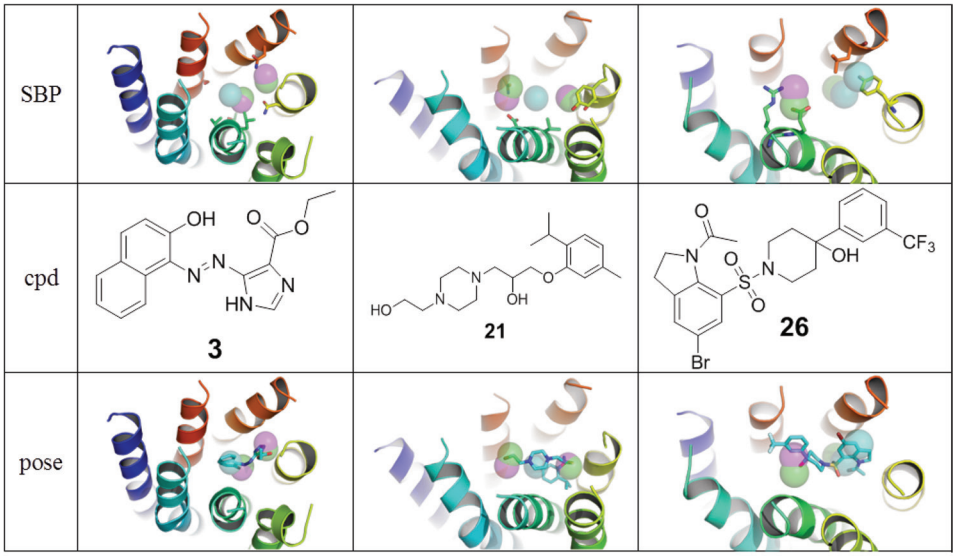


Figure 7.6: The important interacting residues for the different receptors with the pharmacophores and fitted compounds. AA2AR: related to acceptor features: 5.42 & 6.55, 3.36 & 7.42, to donor features: 5.42 & 6.55, 3.36 & 7.42, hydrophobic feature: 3.32 & 3.33. ADRB2: acceptors: 5.38&5.42&5.46, 7.39, donors: 3.32, 5.38&5.42&5.46, hydrophobic: 3.33. S1PR1: acceptors: 3.28, 5.39, donor: 3.29 and hydrophobics: 5.39&6.55, 5.39&6.55.

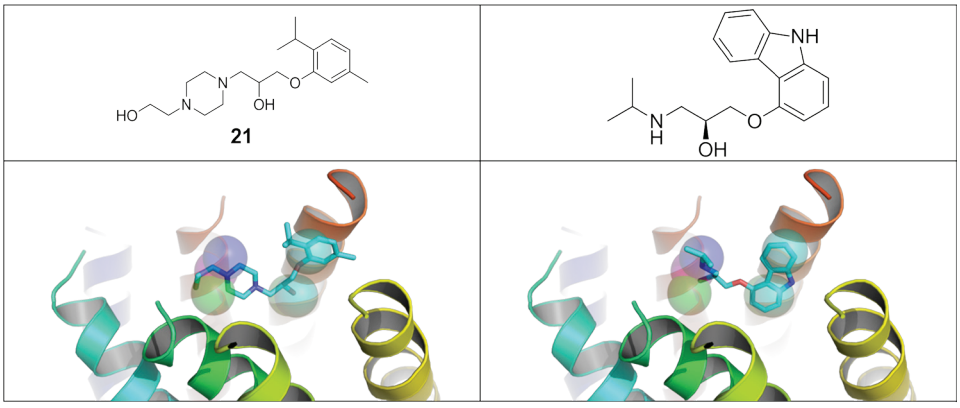


Figure 7.7: Structure-based pharmacophore for S1PR1 with a) compound 21 and b) carazolol in the conformation as found in the crystal structure of ADRB2[38].

7.4 Conclusions

We performed compound selections for three receptors based on a combined ligand- and structure-based approach. Retrospective analysis of both separate methods on the AA2AR and ADRB2 receptor indicated that the results depend on the chemical diversity amongst active compounds for a target. The diverse AA2AR ligand training set facilitated the construction of a robust ligand-based model superior to the protein-based model, while a combination of a ligand- and protein-based model gave the best results for ADRB2. Using a consensus scoring procedure combining the ligand-based and structure-based tools, we performed one of the first large-scale in-vitro prospective chemogenomic screening exercises, experimentally testing all in silico hits selected for all investigated targets. Out of the total of 900 compounds screened, hit rates varied from 2% (at the AA2A receptor), over 0.7% (at the ADRB2 receptor) to 0.3% (at the S1P1 receptor). While these variations in hit rates might not be surprising by themselves, what is certainly surprising is the high degree of cross-reactivity encountered: In several cases, compounds were not found to be active against the receptor they were selected for, but rather against a different member of the panel. As a conclusion of practical relevance, which is supported also by the results obtained in this current work, the authors advocate more 'fuzzy' virtual screening setups to be considered as an alternative to in particular very specifically defined pharmacophores – since as we have seen in this work, also information from bioactive compounds (or receptors) of related proteins may very well be relevant to identify novel bioactive chemical matter. While the screens performed here, generating thousands of bioactivity data points, is relatively large for academic setups, it still considers only a relatively small number of receptors. Ideally, the chemogenomics approach presented here should be extended with more compounds and a larger panel of targets and should preferably be tested using the same assay technology. Such endeavors are already frequently employed for other protein families, like for kinases and will provide more complete information to medicinal and computational chemists to optimize screening hits and leads and design better screening libraries than single-target screens can.[75]

References

1. Kubinyi, H. and G. Muller, eds. Chemogenomics in Drug Discovery: A Medicinal Chemistry Perspective., ed. R. Mannhold and G. Folkers. 2004, Wiley-VCH.
2. Bredel, M. and E. Jacoby, Chemogenomics: an emerging strategy for rapid target and drug discovery. Nat Rev Genet, 2004. **5**(4): p. 262-75.
3. Harris, C.J. and A.P. Stevens, Chemogenomics: structuring the drug discovery process to gene families. Drug Discov Today, 2006. **11**(19-20): p. 880-8.
4. Caron, P.R., et al., Chemogenomic approaches to drug discovery. Current Opinion in Chemical Biology, 2001. **5**(4): p. 464-70.
5. Rognan, D., Chemogenomic approaches to rational drug design. Br J Pharmacol, 2007. **152**(1): p. 38-52.
6. Rognan, D., Structure-Based Approaches to Target Fishing and Ligand Profiling. Molecular Informatics, 2010. **29**(3): p. 176-87.
7. Keiser, M.J., et al., Predicting new molecular targets for known drugs. Nature, 2009. **462**(7270): p. 175-81.
8. Badrinarayan, P. and G.N. Sastry, Virtual High-throughput Screening in New Lead Identification. Comb Chem High Throughput Screen, 2011.
9. Seifert, M.H., J. Kraus, and B. Kramer, Virtual high-throughput screening of molecular databases. Curr Opin Drug Discov Devel, 2007. **10**(3): p. 298-307.
10. Reddy, A.S., et al., Virtual screening in drug discovery -- a computational perspective. Curr Protein Pept Sci, 2007**. 8**(4): p. 329-51.
11. Brianso, F., et al., Cross-pharmacology analysis of G protein-coupled receptors. Curr Top Med Chem, 2011. **11**(15): p. 1956-63.
12. Mestres, J., et al., Data completeness--the Achilles heel of drug-target networks. Nature Biotechnology, 2008. **26**(9): p. 983-4.
13. Overington, J.P., B. Al-Lazikani, and A.L. Hopkins, How many drug targets are there? Nat Rev Drug Discov, 2006. **5**(12): p. 993-6.
14. Knox, C., et al., DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res, 2011. **39**(Database issue): p. D1035-41.
15. Chen, X., Y. Lin, and M.K. Gilson, The binding database: overview and user's guide. Biopolymers, 2001. **61**(2): p. 127-41.
16. Chen, X., et al., The Binding Database: data management and interface design. Bioinformatics, 2002. **18**(1): p. 130-9.
17. Chen, X., M. Liu, and M.K. Gilson, BindingDB: a web-accessible molecular recognition database. Comb Chem High Throughput Screen, 2001. **4**(8): p. 719-25.
18. Wang, R., et al., The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J Med Chem, 2004. **47**(12): p. 2977-80.
19. Wang, R., et al., The PDBbind database: methodologies and updates. J Med Chem, 2005. **48**(12): p. 4111-9.
20. Hu, L., et al., Binding MOAD (Mother Of All Databases). Proteins, 2005. **60**(3): p. 333-40.
21. Smith, R.D., et al., Exploring protein-ligand recognition with Binding MOAD. J Mol Graph Model, 2006. **24**(6): p. 414-25.
22. Oprea, T.I., et al., Lead-like, drug-like or "Pub-like": how different are they? J Comput Aided Mol Des, 2007. **21**(1-3): p. 113-9.
23. Okuno, Y., et al., GLIDA: GPCR--ligand database for chemical genomics drug discovery--database and tools update. Nucleic Acids Res, 2008. **36**(Database issue): p. D907-12.
24. Krejsa, C.M., et al., Predicting ADME properties and side effects: the BioPrint approach. Curr Opin Drug Discov Devel, 2003. **6**(4): p. 470-80.
25. Bender, A., et al., Chemogenomic data analysis: prediction of small-molecule targets and the advent of biological fingerprint. Comb Chem High Throughput Screen, 2007. **10**(8): p. 719-31.
26. van der Horst, E., et al., Substructure mining of GPCR ligands reveals activity-class specific functional groups in an unbiased manner. J Chem Inf Model, 2009. **49**(2): p. 348-60.
27. Klabunde, T., Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. Br J Pharmacol, 2007. **152**(1): p. 5-7.

28. Geppert, H., M. Vogt, and J. Bajorath, Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. J Chem Inf Model, 2010. **50**(2): p. 205-16.
29. Bender, A., et al., How similar are similarity searching methods? A principal component analysis of molecular descriptor space. J Chem Inf Model, 2009. **49**(1): p. 108-19.
30. Leach, A.R., et al., Three-dimensional pharmacophore methods in drug discovery. J Med Chem, 2010. **53**(2): p. 539-58.
31. Moitessier, N., et al., Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. British Journal of Pharmacology, 2008. **153** Suppl 1: p. S7-26.
32. Sanders, M.P.A., et al., From the protein's perspective: The benefits and challenges of protein structure-based pharmacophore modeling. MedChemComm, 2011. submitted.
33. Kolb, P., et al., Docking and chemoinformatic screens for new ligands and targets. Current Opinion in Biotechnology, 2009. **20**(4): p. 429-36.
34. Liu, X., et al., PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. Nucleic Acids Res, 2010. **38**(Web Server issue): p. W609-14.
35. Rask-Andersen, M., M.S. Almen, and H.B. Schioth, Trends in the exploitation of novel drug targets. Nat Rev Drug Discov, 2011. **10**(8): p. 579-90.
36. Szabo, N., Z.T. Kincses, and L. Vecsei, Novel therapy in Parkinson's disease: adenosine A(2A) receptor antagonists. Expert Opin Drug Metab Toxicol, 2011. **7**(4): p. 441-55.
37. Dorsam, R.T. and J.S. Gutkind, G-protein-coupled receptors and cancer. Nat Rev Cancer, 2007. **7**(2): p. 79-94.
38. Cherezov, V., et al., High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science, 2007. **318**(5854): p. 1258-65.
39. Jaakola, V.P., et al., The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. Science, 2008. **322**(5905): p. 1211-7.
40. Horuk, R., Chemokine receptors. Cytokine Growth Factor Rev, 2001. 12(4): p. 313-35.
41. Sanders, M.P., et al., Snooker: A Structure-Based Pharmacophore Generation Tool Applied to Class A GPCRs. J. Chem. Inf. Model, 2011. **51**(9): p. 2277-92.
42. van Loenen, P.B., et al., Agonist-dependent effects of mutations in the sphingosine-1-phosphate type 1 receptor. Eur. J. Pharmacol., 2011. **667**: p. 105 - 112.
43. Gasteiger, J., C. Rudolph, and J. Sadowski, Automatic generation of 3D-atomic coordinates for organic molecules Tetrahedron Comput. Methods, 1990. **3**: p. 537 - 547.
44. Liu, X., et al., Cyndi: a multi-objective evolution algorithm based method for bioactive molecular conformational generation. BMC Bioinformatics, 2009. **10**: p. 101.
45. van der Horst, E., et al., Substructure-Based Virtual Screening for Adenosine A2A Receptor Ligands. ChemMedChem, 2011. (in press).
46. Bender, A., Databases: Compound bioactivities go public. Nat Chem Biol, 2010. **6**: p. 309 - 309.
47. Irwin, J.J. and B.K. Shoichet, ZINC--a free database of commercially available compounds for virtual screening. J Chem Inf Model, 2005. **45**(1): p. 177-82.
48. Kazius, J., et al., Substructure mining using elaborate chemical representation. J Chem Inf Model, 2006. **46**(2): p. 597-605.
49. Nijssen, S. and J.N. Kok. A quickstart in frequent structure mining can make a difference. in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004.
50. Del Rio, A., et al., CoCoCo: a free suite of multiconformational chemical databases for high-throughput virtual screening purposes. Mol Biosyst, 2010. **6**(11): p. 2122-8.
51. Lipinski, C.A., et al., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev, 2001. **46**(1-3): p. 3-26.
52. Wenlock, M.C., et al., A comparison of physiochemical property profiles of development and marketed oral drugs. J Med Chem, 2003. **46**(7): p. 1250-6.
53. Verzijl, D., et al., Noncompetitive antagonism and inverse agonism as mechanism of action of nonpeptidergic antagonists at primate and rodent CXCR3 chemokine receptors. J Pharmacol Exp Ther, 2008. **325**(2): p. 544-55.
54. van der Lee, M.M.C., et al., beta-Arrestin recruitment assay for the identification of agonists of the sphingosine 1-phosphate receptor EDG1. J Biomol Screen, 2008. **13**: p. 986-998.

55. Xu, F., et al., Structure of an agonist-bound human A2A adenosine receptor. Science, 2011. **332**(6027): p. 322-7.
56. Jiang, Q., et al., Hydrophilic side chains in the third and seventh transmembrane helical domains of human A2A adenosine receptors are required for ligand recognition. Mol Pharmacol, 1996. **50**(3): p. 512-21.
57. Kim, J., et al., Site-directed mutagenesis identifies residues involved in ligand recognition in the human A2a adenosine receptor. J Biol Chem, 1995. **270**(23): p. 13987-97.
58. Kim, J., et al., Glutamate residues in the second extracellular loop of the human A2a adenosine receptor are required for ligand recognition. Mol Pharmacol, 1996. **49**(4): p. 683-91.
59. Moro, S., et al., Progress in the pursuit of therapeutic adenosine receptor antagonists. Med Res Rev, 2006. **26**(2): p. 131-59.
60. Xu, F., et al., Structure of an agonist-bound human A2A adenosine receptor. Science. **332**(6027): p. 322-7.
61. Strader, C.D., et al., Conserved aspartic acid residues 79 and 113 of the beta-adrenergic receptor have different roles in receptor function. J Biol Chem, 1988. **263**(21): p. 10267-71.
62. Strader, C.D., et al., Identification of residues required for ligand binding to the beta-adrenergic receptor. Proc Natl Acad Sci U S A, 1987. **84**(13): p. 4384-8.
63. Liapakis, G., et al., The forgotten serine. A critical role for Ser-2035.42 in ligand binding to and activation of the beta 2-adrenergic receptor. J Biol Chem, 2000. **275**(48): p. 37779-88.
64. Suryanarayana, S. and B.K. Kobilka, Amino acid substitutions at position 312 in the seventh hydrophobic segment of the beta 2-adrenergic receptor modify ligand-binding specificity. Mol Pharmacol, 1993. **44**(1): p. 111-4.
65. Strader, C.D., et al., Identification of two serine residues involved in agonist activation of the beta-adrenergic receptor. J Biol Chem, 1989. **264**(23): p. 13572-8.
66. Wieland, K., et al., Involvement of Asn-293 in stereospecific agonist recognition and in activation of the beta 2-adrenergic receptor. Proc Natl Acad Sci U S A, 1996. **93**(17): p. 9276-81.
67. Kikkawa, H., et al., The role of the seventh transmembrane region in high affinity binding of a beta 2-selective agonist TA-2005. Mol Pharmacol, 1998. **53**(1): p. 128-34.
68. Pham, T.C., et al., Molecular recognition in the sphingosine 1-phosphate receptor family. J Mol Graph Model, 2008. **26**(8): p. 1189-201.
69. Parrill, A.L., et al., Identification of Edg1 receptor residues that recognize sphingosine 1-phosphate. J Biol Chem, 2000. **275**(50): p. 39379-84.
70. Wang, D.A., et al., A single amino acid determines lysophospholipid specificity of the S1P1 (EDG1) and LPA1 (EDG2) phospholipid growth factor receptors. J Biol Chem, 2001. **276**(52): p. 49213-20.
71. Surgand, J.S., et al., A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors. Proteins, 2006. **62**(2): p. 509-38.
72. Gozalbes, R., et al., QSAR Strategy and Experimental Validation for the Development of a GPCR Focused Library. QSAR Comb. Sci., 2005. **24**(4): p. 508 - 516.
73. Steffen, A., et al., Comparison of molecular fingerprint methods on the basis of biological profile data. J Chem Inf Model, 2009. **49**(2): p. 338-47.
74. Wawer, M. and J. Bajorath, Similarity-potency trees: a method to search for SAR information in compound data sets and derive SAR rules. J Chem Inf Model, 2010. **50**(8): p. 1395-409.
75. Goldstein, D.M., N.S. Gray, and P.P. Zarrinkar, High-throughput kinase profiling as a platform for drug discovery. Nat Rev Drug Discov, 2008. **7**(5): p. 391-7.

CHAPTER 8

A comparative analysis of pharmacophore screening tools

Marijn P.A. Sanders[1], Armenio J.M. Barbosa[2], Barbara Zarzycka[3], Gerry A.F. Nicolaes[3], Jan P.G. Klomp[4], Jacob de Vlieg[1,5], Alberto Del Rio[2]

[1]CMBI, NCMLS, Radboud University Nijmegen Medical Centre, Nijmegen, the Netherlands; [2]Department of Experimental Pathology, Alma Mater Studiorum, University of Bologna, Bologna, Italy; [3]Department of Biochemistry, Cardiovascular Research Institute Maastricht, Maastricht University, Maastricht, The Netherlands; [4]Lead Pharma Medicine, , Nijmegen, The Netherlands; [5]Netherlands eScience Center, Amsterdam, The Netherlands

Journal of Chemical Information and Modelling, submitted

Acknowledgements

We thank Dave Wood for critical reading the manuscript and all software companies and developers for their support. This work was supported by: AIRC Emilia Romagna Start-up grant 6266 (A.J.M.B. and A.D.); Bayer Haemophilia Awards programme (G.A.F.N); Dutch Top Institute Pharma, project number: D1-105 (M.P.A.S.); Cyttron II - FES0908 (B.Z)

Abstract

The pharmacophore concept is very important in Computer-Aided Drug Design (CADD) mainly due to its application in High-Throughput Virtual Screening (HTVS). With many pharmacophore screening software available, it is of the utmost interest to explore the behavior of these tools when applied to different biological systems. In this work we present a comparative analysis of eight pharmacophore screening algorithms (Catalyst, Unity, LigandScout, Phase, Pharao, MOE, Pharmer and POT) for their use in typical HTVS campaigns against four different biological targets. The results herein presented show how the performance of each pharmacophore screening tool might be specifically related to factors such as the characteristics of the binding pocket, the use of specific pharmacophore features and the use of these techniques in specific steps/contexts of the drug discovery pipeline. We conclude that algorithms with overlay-based scoring functions are generally slower than RMSD-based scoring functions but have a better performance in compound library enrichment. This observation together with other findings can be used to choose the most appropriate algorithm for specific virtual screening projects. We also analyzed how pharmacophore algorithms can be combined together in order to increase the success of hit compound identification. Furthermore, this study provides a valuable benchmark set for further developments in the field of pharmacophore search algorithms e.g. by using pose predictions and compound library enrichment criteria as described in this work.

8.1 Introduction

In the field of drug design, high-throughput virtual screening (HTVS) methods encompass a valuable set of computational approaches for the analysis of large chemical structure libraries with the purpose of identifying hit compounds capable of interacting with a biological target of interest [1]. While combinatorial chemistry and high throughput screening (HTS) procedures over the last few decades have represented an important step in drug discovery to accumulate large amount of data, the global importance of *in silico* techniques is vice versa ascribable to reduced costs and the increased time-efficiency to unveil new potential active compounds [2, 3]. Among all computational approaches that can help to guide drug discovery, the so-called structure-based (SB) design approaches, which use the three-dimensional information of the biological target, are among the most popular [4-6]. However, despite the existence of large numbers of apparently different computational approaches, recent studies emphasize the usage of simple and already established techniques for the successful disclosure of important information towards the selection of novel bioactive compounds [7, 8]. In this context, some seemingly old concepts, such as that of the pharmacophore, have proven to be extremely useful over the past 30 years and it is surprising that many of these concepts are regaining momentum [9-11]. For pharmacophore modelling, in particular, this *renaissance* has also been fostered by the current possibility to generate hypotheses directly from crystallographic, NMR or computational models of protein-ligand complexes [9, 10, 12]. Together with the fact that nowadays three-dimensional structure of biological receptors and enzymes become available much more frequently than in the past, one would theoretically need few steps in order to rapidly setup pharmacophore screening campaigns towards the selection of novel molecular entities for biological testing and/or lead optimization purposes.

Beside the increased availability of structural information on pharmacological targets, advances in computing power and improvements in screening algorithms, presently pharmacophore screenings are also further stimulated by the increasing number of academic services and chemical vendors that offer large databases of commercially available compounds and virtual libraries for this purpose [13-17]. Some of these databases allow the circumvention of typical preparation steps such as hydrogen addition, tautomer and stereoisomers enumeration and, most importantly, conformer generation [17]. From the practical point of view, pharmacophores can be used to screen millions of high quality compounds structures within a reasonable amount of time, particularly when approximations such as a rigid pharmacophore fitting procedure are used. Like molecular docking, pharmacophore search algorithms should not only discriminate between active and inactive compounds but should also correctly orient the ligand in the protein-binding region [18, 19].

While many studies typically focusing on the accurate reproduction of binding modes and compound library enrichment have been published in the last year to assess docking screening algorithms [5, 20, 21], the pharmacophore concept, introduced above, has

been the object of few comparative studies assessing the performance of pharmacophore screening tools [9, 12, 19, 22, 23].

Here we will present an assessment of eight free and commercial software packages for pharmacophore screening. In order to ensure a fair comparison of the screening algorithms, we have chosen to manually curate pharmacophores extracted from X-ray structures, and to perform the virtual screens on rigid compound structures of pre-calculated conformers using default settings. Four biological targets were analyzed, for which the locations of a large number of ligands elucidated by X-ray crystallography were collected from the literature, and were used to derive structure-based pharmacophores and evaluate pharmacophore screening performance.

The accuracy of the predicted binding modes as well as library enrichments for the different pharmacophore search algorithms is analyzed to elucidate how different factors, e.g. pharmacophore hypotheses and conformational states, may influence the outcome of high-throughput screenings.

8.2 Methods

Figure 8.1 shows the computational protocol that was applied to each biological target. Full details of each step are discussed in the next paragraphs.

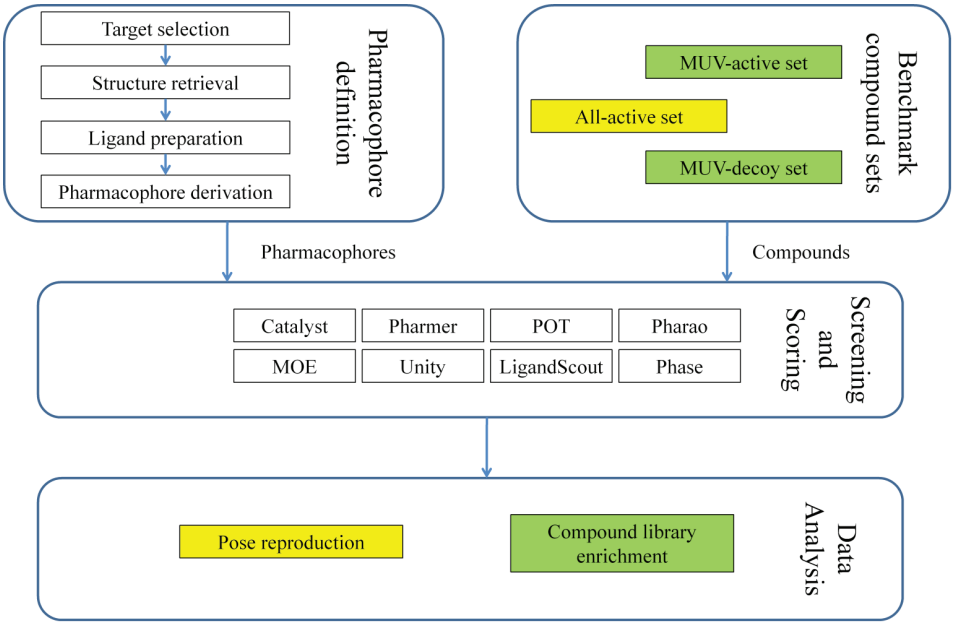


Figure 8.1: Flow diagram of the dataset preparation protocol and computational procedures. Yellow boxes indicate procedures in which all co-crystallized ligands are used, green boxes indicate procedures where the unbiased active and decoy set is used.

8.2.1. Pharmacophore definition

8.2.1.1. Target selection

Four datasets corresponding to different biological targets were used in this study, namely CDK2 (Cyclin-dependent kinase 2), Chk-1 (Checkpoint kinase 1), PTP-1B (Protein tyrosine phosphatase 1B) and Urokinase. The last three protein targets were chosen in order to take advantage of published data from Brown et al. [24] while the CDK2 dataset was created from crystal structures retrieved from the Protein Data Bank [25]. All the biological targets have been selected for their important roles in different biological processes. CDK2 is a well-known protein kinase involved in the control of the cell cycle [26]; Chk-1 is a kinase required for checkpoint mediated cell cycle arrest in response to DNA damage or the presence of unreplicated DNA [27]; PTP-1B, is a regulator involved in insulin signaling and has been implicated as a potential therapeutic target for treatment of type II diabetes [28]; Urokinase, is a serine protease that circulates in plasma and that has been implicated in a number of tumor-related activities [29].

8.2.1.2. Structure retrieval

Ligand coordinates for PTP1B, CHK1 and Urokinase datasets were obtained from Brown et al. [24] and were aligned with the PDB IDs 1NZ7, 2YWP and 1OWK, respectively. CDK2 was selected because of the high number of complexes with different co-crystallized ligands available for the biological target in the Protein Data Bank [25, 26]. Reference X-ray structures were obtained from the PDB (accessed November 2010)[25] by using the human CDK2 Uniprot [30] Accession ID: P24941. The collected CDK2 complexes were filtered in order to retrieve proteins with bound ligands and no modified residues resulting in a set of 107 CDK2 complexes. These complexes were visually inspected and discarded in cases where ligand atoms were missing or multiple ligand conformations were present in a single PDB entry. Any duplicate entries were removed which resulted in a non-redundant set of 80 complexes (a full list is available as supplementary material). Finally, all complexes were superimposed with Sybyl [31] and the ligands were extracted.

8.2.1.3. Ligand preparation

After manual correction of ligands for all targets (e.g. bond orders, aromaticity and charges), Epik software [32] was used to calculate physiological protonation states. A visual inspection of the ligand embedded in the original protein PDB structure was performed to check for correct tautomeric forms.

8.2.1.4. Pharmacophore derivation

Ligand structures were given appropriate atoms-types using an *in-house* rule-based classification system developed at Organon NV and pharmacophore features were generated after application of Renner's fuzzy pharmacophore algorithm with an Rc-value of 2.0Å [33]. Manual inspection and combination of features shared by most ligands

resulted in the final pharmacophores that are depicted in **section 8.3**. Excluded volume features were added by an iterative procedure. Starting with the addition of an excluded volume feature at the position of the closest atom in the protein to a co-crystallized ligand atom, the algorithm continues to add excluded volumes until all protein atoms are considered. This addition is recursively performed for atom distances between 3.0Å -6.0Å from the co-crystallized ligands and that are at least 1.0Å from the nearest excluded volume feature. In contrast to most algorithms, which use the atom centre to evaluate if a pose is inside the excluded volume, Phase rejects ligand poses with an overlap of the ligand VDW radius with the excluded volumes. To be more consistent with the other used algorithms we therefore reduced the excluded volume radii of Phase pharmacophores with 1.7Å (approximately the VDW radius of a Carbon atom).

8.2.2. Benchmark compound sets

8.2.2.1. All-actives set

The compound sets comprising all active compounds of each respective target (**Table 8.1**) were used to assess how well each individual algorithm performs in the reproduction of the crystal structure pose by means of RMSD calculation.

8.2.2.2. Maximum unbiased validation sets of active and decoys

In order to assess the compound library enrichment of the different pharmacophore screens, data sets of actives and decoys were designed so as to avoid analogue bias (overrepresentation of certain scaffolds or chemical entities), and artificial enrichment (classification is caused by differences in simple physicochemical descriptors like molecular weight, number of bonds and acceptors, donors, rather than correct representation of the protein-ligand interactions). First, for each target, a set of assumed inactives compounds was prepared from the CoCoCo database [14, 17] by selecting compounds with a BCI-Tanimoto fingerprint similarity [34] of 0.5 or less to at least one ligand reported in the ChEMBL [35, 36] for that particular target. Second, the maximum unbiased validation (MUV) sets protocol [37] was used to select 30 actives and 15,000 decoys (consistent with the set sizes available on the MUV website [38]) per biological target from each subset of the CoCoCo database. This protocol ensures an unbiased validation set by maximizing “active-active distances” G(t) and “active-decoy distances” F(t). Numerical integration of both distribution functions enables computation of global figures for data set self-similarity (ΣG) and the separation between active and decoys (ΣF). A parameter describing the “data set clumping” S(t), and its numerical integral ΣS can be calculated by subtraction of G(t) from F(t). Negative values of ΣS indicate clumping of actives, while positive values indicate dispersion of actives and clumping of small clusters of decoys with single active compounds, and values near zero indicate a spatially random distribution of actives and decoys [37]. The data sets generated with the MUV protocol will be used for the compound library enrichment studies and will be referred in the text

as MUV-active or MUV-decoys.

For all compound data sets (All-actives, MUV-actives and MUV-decoys) of the four biological targets, three-dimensional conformations were generated with the Confgen software using the *comprehensive* algorithm [39].

Table 8.1: Overview of datasets included in the present study. All files in single- and multi-conformation are available in supplementary materials.

Target	All-actives	All-actives Conformations	MUV-decoys	MUV-decoys Conformations	MUV-actives	MUV-actives Conformations
CDK2	80	992	15000	248250	30	352
CHK1	123	1913	15000	287203	30	457
PTP1B	110	4634	15000	405398	30	1123
Urokinase	75	703	15000	268646	30	192

8.2.3. Screening and Scoring

8.2.3.1. Screening

Screens were performed using eight different software tools as shown in **Table 8.2** below. To mimic the scenario in which non-experts might apply pharmacophore screens and in order to avoid artificial bias towards certain algorithms, we ran all algorithms with default parameters with the exception of Pharao and Pharmer. For Pharao we used the *tversky_ref* score instead of the Tanimoto score as the paper indicated that this score was most suitable for scenarios where it is important that as many features of the pharmacophore are matched as possible. Since Pharmer cannot handle excluded volume features, we post-processed the Pharmer poses with POT with use of only the excluded volume definition and without a fitting. As the virtual screens were performed in different labs and using different hardware and software systems we decided to focus entirely on the quality of the produced results by the different algorithms and to disregard the CPU-timing which is required for the presented pharmacophore searches.

8.2.3.1. Scoring

Internal molecular symmetries were considered by calculation of the Root Mean Square Deviations (RMSD) of all possible structural matches of fitted conformations and corresponding co-crystallized reference poses using RDkit [40] functionality. The lowest RMSD value, corresponding to the best fit, was reported for each pose and used for all further analysis. Receiver Operator Characteristic (ROC) curves, Area Under the ROC Curves (AUC) and enrichment factors were calculated after ranking compounds from the MUV-active and MUV-decoys set based on the score values as reported in **Table 8.2**. Enrichment factors (*EF*) after *x*% of the library screened were calculated according to the following formula ($N_{experimental}$ = number of experimentally found active structures in the top *x*% of the sorted database, $N_{expected}$ = number of expected active structures, N_{active} = total number of active structures in database).

$$EF = \frac{N^{x\%}_{experimental}}{N^{x\%}_{expected}} = \frac{N^{x\%}_{experimental}}{N_{active} \cdot x\%} \quad (1)$$

Table 8.2: List of pharmacophore screening algorithms used in this study.

Tool	Version	Scoring algorithm[a]	Scoring method[b]	Best[c]	Tool availability	Reference
Catalyst	Discovery Studio v2.5.5.9350	FitValue [d]	Overlay	High	Commercial	[41]
Pharmer	-	RMSD	RMSD	Low	Open-source	[42]
POT	-	RMSD	RMSD	Low	Open-source[e]	[43]
Pharao	3.0.3	Tversky_ref[f]	Overlay[f]	High	Open-source	[44]
MOE	2010.10 (date)	RMSD	RMSD	Low	Commercial	[45]
Unity	Sybyl X1.1.1	QFIT[g]	Overlay	High	Commercial	[31]
LigandScout	3.02	Pharmacophore Fit[h]	Overlay	High	Commercial	[46, 47]
Phase	3.3	Fitness[i]	RMSD[i]	High	Commercial	[48]

[a] Scoring algorithm refers to the tool routine that was used in the present study to score ligand poses.
[b] Scoring method refers to the methodology class of scoring algorithm. RMSD-based methods check the distance of the feature group of the compound to the pharmacophore feature center. Overlay methods take the radii of the features and/or atoms into account and use this to assess how well a feature is matched.
[c] Best refers to the score value (high/low) which is used to select the best match.
[d] FitValue evaluates for each molecular feature the distance to the center of the pharmacophore feature with respect to the pharmacophore feature radius.
[e] Will be made available soon.
[f] Tversky_ref evaluates the volume-overlap of features of the pharmacophore and compound with respect to the volume of the pharmacophore. Pharao uses Gaussian overlaps of pharmacophore features and ligand atoms and does not require that all features are matched. Poses with a lower number of matched features are able to get as high scores as poses with more matched features.
[g] QFIT is intended to compare alternate mappings of a single compound to the query and choose the best match. While Unity reports hits as soon as they fulfill the query, post processing options to relax and tighten hits to more closely match the query are available within Sybyl package.
[h] Pharmacophore Fit is a simple geometric scoring function that takes into account only chemical feature overlap. Other scoring functions are available within LigandScout package.
[i] Fitness score of Phase is based on an RMSD term, vector term and a term describing the overlay of the produced pose with a reference pose. No vector features and no reference pose were considered in the present study, thus the resulting score is purely RMSD-based.

8.2.4. Data analysis

8.2.4.1. Pose reproduction

Pharmacophoric poses were generated by application of the 8 screening algorithms to the 4 datasets, and the results were analyzed with respect to the accuracy of experimental binding mode reproduction and compound library enrichments. For each molecule, both the pose with the lowest RMSD to the reference structure and the RMSD of the pose with the best score were reported. The cumulative percentage of poses below a certain RMSD was calculated for both X-ray structures and the multi-conformational datasets of actives previously generated. For CHK1, PTP1B and Urokinase multiple pharmacophores are defined which recognize different subsets of active molecules. To evaluate the combined performance for CHK1, PTP1B and Urokinase we also merged the outcomes of the individual pharmacophores searches for these targets. To evaluate the performance of each algorithm in respect of compound library enrichment, the percentage of retrieved actives (sensitivity) versus the percentage of retrieved decoys (specificity) was calculated using the ranking of MUV data sets (actives and decoys) as deduced from their respective score values and visualized in receiver operator curves (ROC). Corresponding area under the ROC curves (AUC) and enrichment values at 0.5%, 1.0%, 2.0% and 5.0% are calculated and reported.

8.2.4.2. Compound library enrichment

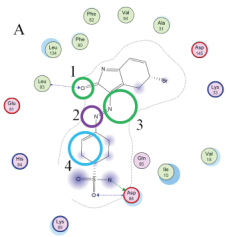
To assess the cooperative behavior of different algorithms we calculated the improvement in enrichment factors for each pharmacophore algorithm when combined with another algorithm. This is calculated as the enrichment of the set of compounds which match in algorithm XY (Y after X) divided by the enrichment of the set of compounds which match in algorithm Y. Such analysis can be useful to identify the combination of algorithms that results in the best enrichment, but also might suggest how fast algorithms can be used as a pre-filter for the slower, but more accurate, algorithms. The heatmaps generated from this analysis (**paragraph 8.3**) show not necessarily which algorithms have the best enrichment, but illustrate the gain in performance achieved by pre-screening with another algorithm. In particular, values below 1.0 indicate that pre-screening with another algorithm will result in worse enrichment, while values above 1.0 indicate that pre-screening is beneficial for the final enrichment value. A value of 1.0 denotes no influence of the second algorithm.

8.3 Results and discussion

8.3.1. CDK2 dataset

8.3.1.1. Pharmacophore perception.

CDK2 is a protein kinase whose pharmacophore features, delineating ligands that target the ATP-binding site, are well described in the literature [49]. Ligand sites typically include a hydrogen bond donor (HBD) and a hydrogen bond acceptor (HBA) representing a pair of key intermolecular interactions occurring with the hinge backbone (feature 1 and 2, **Figure 8.2A**). The importance of these features is exemplified by the fact that almost all the active compounds match those features (**Figure 8.2B**). A second HBA feature is common to most of the active compounds and represents interaction with the gatekeeper residue Glu81 or for bridging water molecules with catalytic Lys33. The hydrophobic feature labeled 4 in **Figure 8.2A** usually matches halogen-substituted aromatic rings that occupy the hydrophobic pocket of the CDK2 ATP binding site.



B

Feature	Matching compounds
1: Acceptor	76
2: Donor	67
3: Acceptor	62
4: Hydrophobe	64
Pharmacophore 1: (1+2+3+4)	45
Number of excluded volumes	130

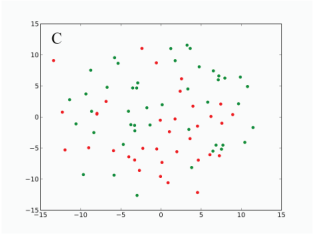


Figure 8.2: CDK2 dataset. **A:** pharmacophore depiction as used in this study on top of PDB entry: 1FVT (note that 1FVT with its co-crystallized ligand is used as a reference and does not contain the donor feature which is present in most ligands co-crystallized with CDK2). **B:** list of pharmacophore features with corresponding matching compounds in the set of actives. **C:** Two dimensional illustration of active compound similarities created using stochastic proximity embedding (SPE) with euclidean distances of BCI fingerprints [50-52]. Green dots represent compounds that match the pharmacophore according to the observed ligand alignment in the crystal structures; red dots are the compounds that do not match the pharmacophore.

8.3.1.2. Retrospective compound set analysis

The 45 compounds statisfying all pharmacophore constraints, in the conformations and positions observed in the crystal structure, are scattered over the chemical space represented by the 80 compounds included in the reference set of actives (**Figure 8.2C**). Most compounds match three of the four features including the acceptor and donor features required for the hydrogenbonding to the *hinge region* of the CDK2.

8.3.1.3. Prospective binding mode reproduction.

The percentage of compounds for which a binding mode is predicted exceeds the 45 compounds (56%) that were found to fulfill the pharmacophore criteria for all algorithms except Catalyst, LigandScout and Unity. Scoring seems to be problematic for most algorithms as only ~20-40% of compounds have a RMSD to the co-crystalized reference pose below 3.0Å for the best scored pose (**Figure 8.3A**). This behaviour appears to be partly caused by the number of conformers that are used as inputs for benchmarking the different algorithms, as performance is better if the best scored pose amongst the X-ray conformation active set is evaluated (**Fig 8.3B,D**).

An analysis of the best reported RMSD of all matched poses for each molecule reveals that ~40% of compounds are predicted with a RMSD below 2.5Å, with the majority of those even below 2.0Å (**Figure 8.3C**). MOE even reaches almost 56% percent. Again, we find that pose prediction is more accurate if only the co-crystalized conformation of the ligand is used for the pharmacophore search (**Fig 8.3D**). However, the increase of compounds predicted below RMSD 2.5Å is only minor (**Figure 8.3C,D**), indicating that the multiple conformation generation protocol used includes at least one conformation representative for the co-crystalized conformer.

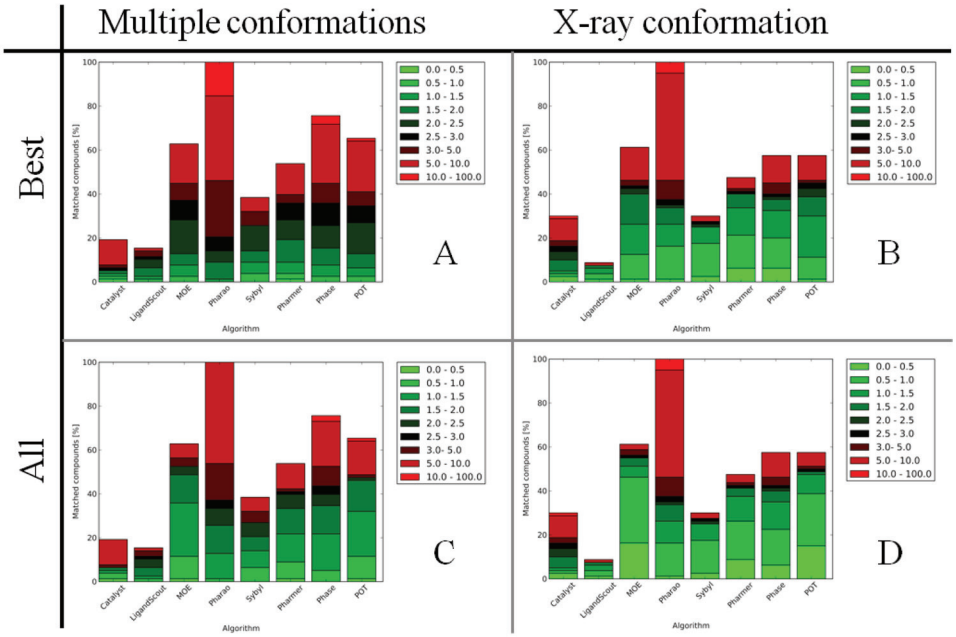


Figure 8.3: RMSD ranges of matched compounds from the co-crystallized ligand in four different scenarios. **A:** the best ranked pose from the ligand set in their multi-conformational format; **B:** the best ranked pose from the ligand set in X-ray conformation; **C:** lowest RMSD from all poses from the ligand set in multi-conformational format. **D:** lowest RMSD from all poses in the ligand set in their X-ray conformation.

8.3.3.4. Compound library enrichment.

Receiver Operator Characteristic curves (ROC) are generated from the results of the pharmacophore searches of the MUV-actives and MUV-decoys sets. Most algorithms retrieve less then 70% and 50% of actives and decoys, respectively, as indicated by the endpoints of the lines in **Figure 8.4A**. The AUC calculation returns values between 0.5 and 0.6 for most algorithms, indicating that the overall enrichment is only slightly better than could be expected from a random selection (**Figure 8.4A,B**). This seems to be mainly because of the relatively low number of active compounds retrieved by the pharmacophore. In particular, LigandScout retrieves only 3 out of 30 actives by matching those actives amongst the first 0.5% of the screened database and reaching an enrichment factor (EF) of 13.33 with AUC of 0.527 (**Figure 8.4B**). Most algorithms show decreasing enrichment factors at higher false positive retrieval rates, indicating that the scoring measures used to rank the compounds that match a pharmacophore are relatively succesful for CDK2 ligand in a compound library enrichment experiment (**Figure 8.4A,B**). The analysis of algorithm combinations (**Figure 8.4C**) shows that LigandScout, Unity and POT are capable of improving enrichment of other algorithms. In particular, LigandScout and POT seem to be complementary as there is an improvement of both

enrichment factors if they are used in a consecutive screening pipeline. Without any prescreening, LigandScout matches 3 out of 30 actives and 711 out of 15,000 decoys resulting in an enrichment factor of 2.1, while POT matches 20 out of 30 actives and 7571 out of 15,000 decoys resulting in an enrichment factor of 1.3. When combined, these algorithms retrieve 3 actives and 471 decoys and have an enrichment factor of 3.2. This is an improvement in enrichment of a factor 1.5 for LigandScout and 2.4 for POT (**Figure 8.4C**). Only few values below 1.0 are observed, indicating that most algorithms can be combined with others without adversely affecting enrichments. This is an important observation that suggests the use of fast algorithms as pre-filtering steps for large compound collections before more accurate and computationally expensive algorithms.

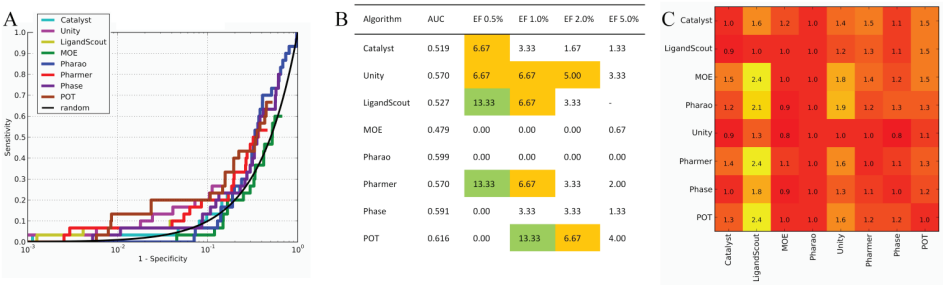


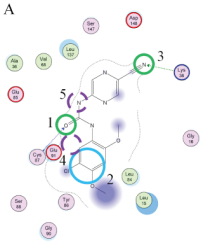
Figure 8.4: Enrichment analysis of CDK2 MUV-dataset. **A:** ROC curves showing the enrichment of CDK2 actives/decoys (dataset created with MUV, see **Table 8.1**). **B:** AUC and enrichment values at 0.5%, 1.0%, 2.0% and 5.0% false positive rate; green indicates an enrichment factor (EF) above 10.0 and orange indicates EF above 5.0. **C:** Heatmap showing the improvement in enrichment factor of the algorithms on the Y-axis if pre-screening with the algorithm on the X-axis is performed.

8.3.4. CHK1 dataset

8.3.4.1. Pharmacophore perception

The pharmacophore features of protein kinase CHK1 are well described in the literature. [53] Similarly to CDK2, a hydrogen bond donor and acceptor pair represents the key interactions for binding the *hinge region* of the kinase (features 1, 4 and 5 of **Figure 8.5A**). However, in this case, two locations of the HBD feature are possible for CHK1 ligands and the analysis of matching compounds indicates their exclusive behaviors (**Figure 8.5B,C**). Other differences with respect to the CDK2 pharmacophore are that the hydrophobic feature (2) (**Figure 8.5A**) is located in the solvent-exposed region of the binding site and is conserved for almost all ligands of the data set (**Figure 8.5B**), and that an additional HBA that is located deep in the binding pocket that represents interactions with the catalytic lysine of the CHK1.

A



B

Feature	Matching compounds
1: Acceptor	120
2: Hydrophobe	110
3: Acceptor	77
4: Donor	48
5: Donor	44
Pharmacophore 1: (1+2+3+4)	36
Pharmacophore 2: (1+2+3+5)	21
Number of excluded volumes	75

C

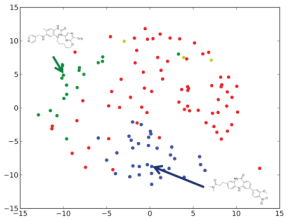


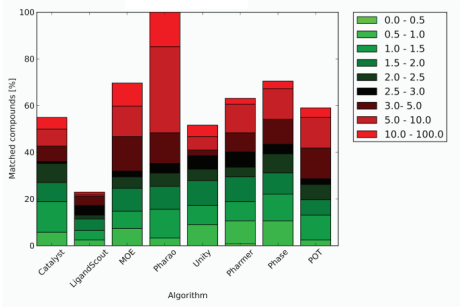
Figure 8.5: CHK1 dataset. **A:** pharmacophore depiction as used in this study on top of PDB entry: 2YWP. Features used in either pharmacophore 1 or 2 are visualized with dashed lines. **B:** list of pharmacophore features with corresponding matching compounds in the set of actives. **C:** Two dimensional illustration of active compound similarities created using stochastic proximity embedding (SPE) with euclidean distances of BCI fingerprints [50-52]. Blue dots represent compounds that match pharmacophore 1; green dots represent compounds that match pharmacophore 2; and yellow dots represent compounds that match both pharmacophores according to the observed ligand alignment in the crystal structures; red dots are the compounds not satisfying the pharmacophore.

8.3.4.2. Retrospective compound set analysis.

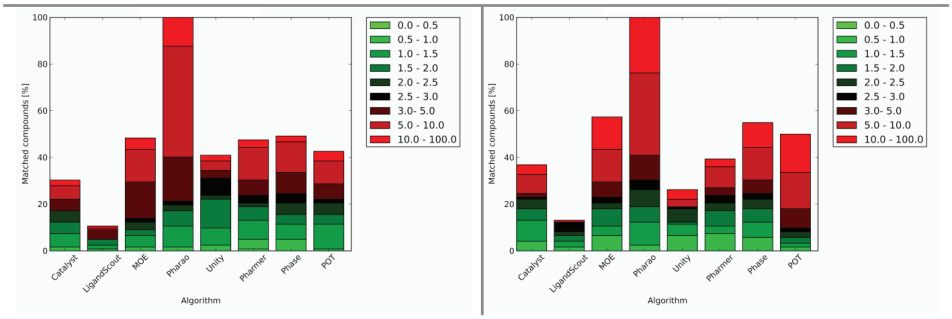
For the reasons provided above, two different pharmacophores were defined for this CHK1 dataset differing in the position of the donor feature (**Figure 8.5A**). As seen in **Figure 8.5C**, the pharmacophores correspond to two topologically distinct clusters of compounds. Among the whole dataset of 123 actives, only three compounds match all five pharmacophore features that are common to pharmacophore 1 and 2 (**Fig 8.5C**). Remarkably those compounds have a relatively high dissimilarity to compounds from both clusters. Most compounds, however, do not match either of the pharmacophores (**Figure 8.5C**, red dots).

8.3.4.3. Prospective binding mode reproduction.

The searches with pharmacophores 1 and 2 retrieve an approximately equal number of compounds and match ~20% of compounds with a RMSD of below 3.0Å (~25 compounds) with respect to the co-crystallized reference structure (**Fig 8.6B,C**). This is in agreement with the retrospective analysis in which pharmacophore 1 and pharmacophore 2 are both derived from ~20% of the compounds. Notably, LigandScout retrieves a low number of actives, while the percentage of compounds in the correct conformations is equal to, or even better than those generated by other algorithms.



Pharmacophore 1+2

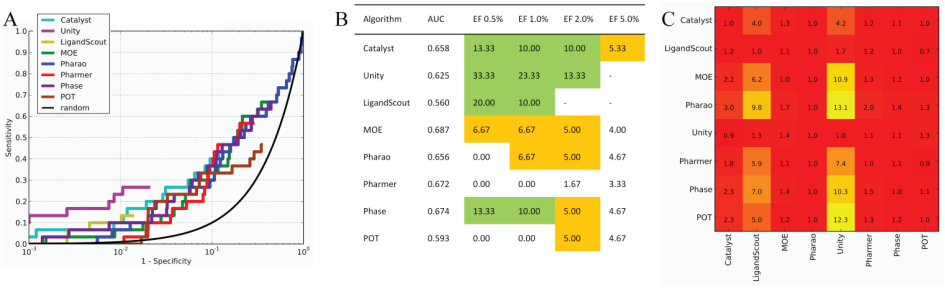


Pharmacophore 1 Pharmacophore 2

Figure 8.6: RMSD ranges of matched compounds from the co-crystallized ligand for pharmacophore 1 and 2; **A:** pharmacophore 1; **B:** and pharmacophore 2 ; **C:**. Both figures refer to the best matching pose of multiconformational data sets.

8.3.4.4. Compound library enrichment.

The stricter matching criteria of LigandScout is reflected in the search of the decoys as just over 1% of compounds are retrieved, while the same analysis shows at least 10% for all other algorithms, except Unity (**Figure 8.7**). These tighter criteria may explain the improved early enrichment of Unity and LigandScout, for which enrichment factors exceeded 20.0 in the top 0.5% of the ranked database (**Figure 8.7B**). RMSD-based scoring methods, like POT, Pharmer, MOE and Phase have also relatively good (≥5.0) enrichments at 2.0% percent of the searched database but do not achieve enrichment factors (EF) of 10.0 or higher. The consecutive screening of compounds with MOE and Unity results in the best enrichment (data not shown), mainly due to the strong performance of Unity which retrieves 8 out of 30 actives and only 306 out of 15,000 decoys. The consecutive application of both algorithms results in an enrichment of 18.7 which is 1.4 times the enrichment factor of the Unity search and 10.9 times the enrichment factor of MOE search. (**Figure 8.7C**).



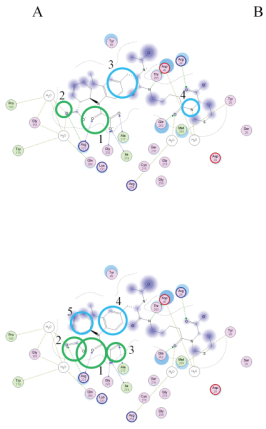
Algorithm	AUC	EF 0.5%	EF 1.0%	EF 2.0%	EF 5.0%
Catalyst	0.658	13.33	10.00	10.00	5.33
Unity	0.625	33.33	23.33	13.33	-
LigandScout	0.560	20.00	10.00	-	-
MOE	0.687	6.67	6.67	5.00	4.00
Pharao	0.656	0.00	6.67	5.00	4.67
Pharmer	0.672	0.00	0.00	1.67	3.33
Phase	0.674	13.33	10.00	5.00	4.67
POT	0.593	0.00	0.00	5.00	4.67

Figure 8.7: Enrichment analysis of CHK1 MUV-dataset. **A**: ROC curves showing the enrichment of CHK1 actives/decoys (dataset created with MUV, see **Table 8.1**). **B:** AUC values and enrichment values at 0.5%, 1.0%, 2.0% and 5.0% false positive rate; green indicates an enrichment factor (EF) above 10.0 and orange indicates EF above 5.0. **C:** Heatmap showing the improvement in enrichment factor of the algorithms on the Y-axis if pre-screening with the algorithm on the X-axis is performed.

8.3.5. PTP1B dataset

8.3.5.1. Pharmacophore perception

PTP1B is a protein tyrosine phosphatase and is characterized by a highly conserved and positively charged active-site [54]. The core of the binding features is characterized by a dyad of hydrogen bond acceptors (HBAs), often represented by acid moieties, that ensure several interactions with arginine and hystidine residues present in the binding pocket. Like CHK1, additional features include a further HBA (pharmacophore 2, feature 3, **Figure 8.8A**) and hydrophobic sites that occupy the binding site region at different locations.



B

Feature	Matching compounds
1: Acceptor	29
2: Acceptor	29
3: Hydrophobe	29
4: Hydrophobe	21
Pharmacophore 1: (1+2+3+4)	21
Number of excluded volumes	62
1: Acceptor	80
2: Acceptor	79
3: Acceptor	78
4: Hydrophobe	80
5: Hydrophobe	77
Pharmacophore 2: (1+2+3+4+5)	77
Number of excluded volumes	62

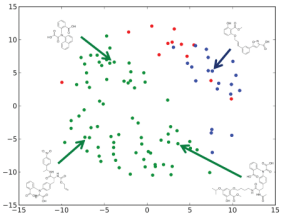


Figure 8.8: PTP1B dataset. **A:** pharmacophore depiction as used in this study on top of PDB entry: 1NZ7. **B:** list of pharmacophore features with the number of matching compounds in the set of actives. **C:** Two dimensional depiction of active compound similarities created using stochastic proximity embedding (SPE) with euclidean distances of BCI fingerprints [50-52]. Blue dots represent compounds that match pharmacophore 1; green dots represent compounds that match pharmacophore 2 according to the observed ligand alignment in the crystal structures; red dots are the compounds not satisfying the pharmacophore.

8.3.5.2. Retrospective compound compound set analysis.

Like CHK1, we defined two distinct pharmacophores for PTP1B (**Figure 8.8A**). These pharmacophores relate to compounds from distinct chemical moieties. For instance, pharmacophore 2, which contains 5 features, matches 77 compounds that vary in size but all contain the N-substituted oxamic acid moiety, while pharmacophore 1 relates only to 21 compounds (**Figure 8.8C**). 12 compounds match neither pharmacophore 1 nor 2.

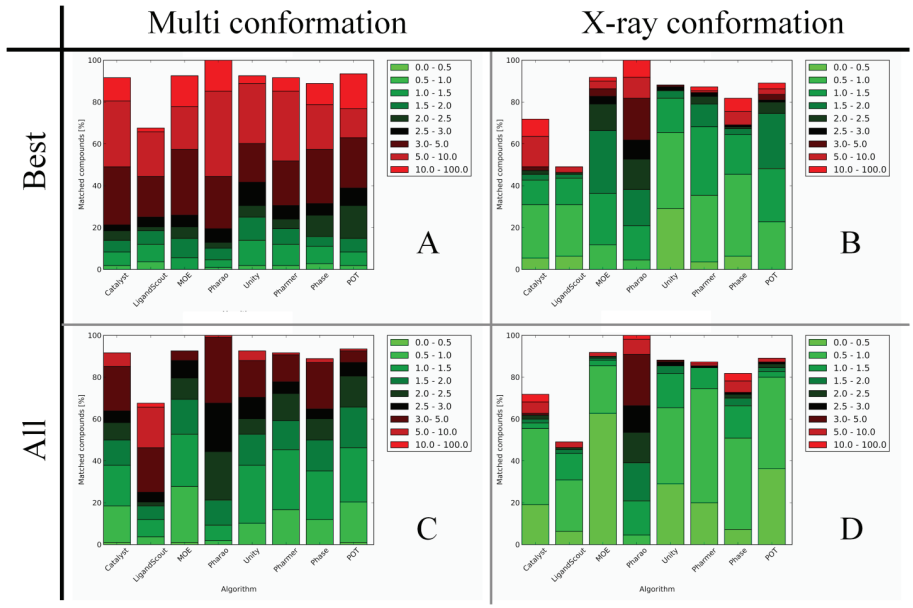


Figure 8.9: RMSD ranges of matched compounds from the co-crystallized ligand in two different scenarios: **A:** the best ranked pose from the ligand set in their multi-conformational format; **B:** the best ranked pose from the ligand set in X-ray conformation; **C:** lowest RMSD from all poses from the ligand set in multi-conformational format. **D:** lowest RMSD from all poses in the ligand set in their X-ray conformation.

8.3.5.3. Prospective binding mode reproduction.

The pharmacophore search based on X-ray conformations (**Figure 8.9B**) shows that 22 compounds (~20%) are retrieved by pharmacophore 1 with an RMSD ≤ 2.5Å and ~70% (77 compounds) by pharmacophore 2. Notably, MOE is able to retrieve ~50% (55 compounds) of the compounds with an RMSD < 2.5Å with a search of pharmacophore definition 1. This is most likely the result of a less stringent feature definition that also explains the relatively high number of retrieved decoys and the moderate performance of MOE in compound library enrichment across all targets (**Fig 8.10A,B**). Conformation generation and appropriate scoring seems, however, to be a problem for nearly all algorithms. Only 20-30% of compounds are matched with an RMSD < 2.5Å (**Figure 8.9A**), while many more are correctly positioned in cases where only the X-ray conformation is used (**Figure 8.9B**) or where the pose with the lowest RMSD is picked amongst all matched poses (**Figure 8.9C**).

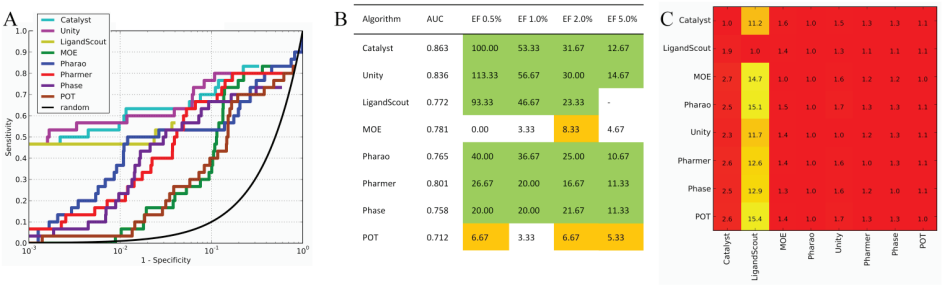


Figure 8.10: Enrichment analysis of PTP1B MUV-dataset. **A:** ROC curves showing the enrichment of PTP1B actives/decoys (dataset created with MUV, see **Table 8.1**). **B:** AUC values and enrichment values at 0.5%, 1.0%, 2.0% and 5.0% false positive rate; green indicates an enrichment factor (EF) above 10.0 and orange indicates EF above 5.0. **C:** Heatmap showing the improvement in enrichment factor of the algorithms on the Y-axis if pre-screening with the algorithm on the X-axis is performed.

8.3.5.4. Compound library enrichment.

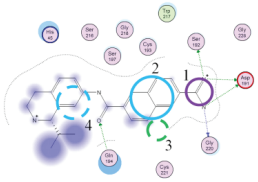
Overlay-based scoring functions (Catalyst, LigandScout and Unity) seem to perform extremely well with respect to compound library enrichment and exhibit very high early enrichment values (**Figure 8.10A,B**). Nonetheless, one should keep in mind that, for all algorithms, scoring is often based on predicted binding modes that do not match the true binding modes. For example, in the cases of Catalyst, LigandScout and to a lesser extent Unity, ~45% of active compounds are ranked above all decoys (**Figure 8.10A**), while only ~20-30% of compounds have RMSD values < 2.5Å (**Figure 8.9A**). A possible explanation for this discrepancy might be the fact that only a fraction of the ligand is represented by pharmacophore features, especially in case of pharmacophore 2 (**Figure 8.8A**). Yet, enrichments are extremely good, especially for overlay-based (Catalyst, LigandScout and Unity) scoring functions (**Figure 8.10B**) and the output of nearly every algorithm can be improved by pre-screening with another algorithm as indicated by the values above 1.0 in **Figure 8.10C**. Combinations with Catalyst, LigandScout, Unity and MOE show the largest enrichment improvement factors (on average > 1.4) in **Figure 8.10C**. The best enrichment factor results from a combination of Catalyst with LigandScout which together retrieve 17 out of 30 actives and 308 out of 15,000 decoys resulting in an enrichment factor of 27.6.

8.3.6. Urokinase dataset

8.3.6.1. Pharmacophore perception.

Despite its name, Urokinase is a serine protease that is clinically used for therapy of thrombolytic disorders and whose small-molecule inhibitors have already been shown to inhibit cancer growth [55]. The pharmacophore created from the list of active compounds (**Figure 8.11A**) shows that two features are always present, specifically a hydrogen bond donor (HBD) (feature 1) and a hydrophobic feature (feature 2). The rest of the pharmacophore consists of one hydrogen bond acceptor (HBA) and another hydrophobic feature that seem to be mutually exclusive since matching compounds are almost complementary (**Figure 8.11B,C**).

A



B

Feature	Matching compounds
1: Donor	74
2: Hydrophobe	74
3: Acceptor	37
4: Hydrophobe	32
Pharmacophore 1: (1+2+3+4)	7
Pharmacophore 2: (1+2+3)	23
Pharmacophore 3: (1+2+4)	36
Number of excluded volumes	54

C

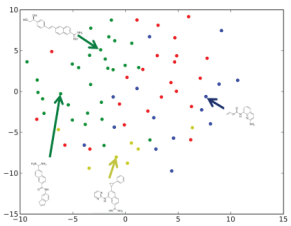


Figure 8.11: Urokinase dataset. **A:** pharmacophore depiction as used in this study on top of PDB entry: 1OWK. Features used in either pharmacophore 2 or 3 are visualized with dashed lines. **B:** list of pharmacophore features with corresponding matching compounds in the set of actives. **C:** Two dimensional illustration of active compound similarities created using stochastic proximity embedding (SPE) with euclidean distances of BCI fingerprints.[50-52] Yellow dots represent compounds that match in pharmacophore 1, 2 and 3; blue dots represent compounds that match pharmacophore 2; green dots represent compounds that match pharmacophore 3 according to the observed ligand alignment in the crystal structures; red dots are the compounds not satisfying the pharmacophore.

8.3.6.2. Retrospective compound set analysis.

Urokinase compounds show relatively diverse features with respect to the other datasets. This is exemplified by the fact that the best four-feature pharmacophore (pharmacophore 1) only satisfied seven compounds (**Figure 8.11**), as deduced from the ligand overlay of co-crystallized ligands. For this reason, we generated two additional three feature pharmacophores which both comprise of three of the features from the original four-feature pharmacophore (**Fig 8.11.A,B**). Pharmacophore 2 contains a donor, hydrophobic and acceptor feature and matched 23 compounds that show little similarity (**Figure 8.11C**). Pharmacophore 3 contains a donor and two hydrophobic features and matches 36 compounds in the co-crystallized overlay. Those compounds are more similar to each other than the compounds matching pharmacophore 2 and are clustered together in topological structure space (**Figure 8.11C**).

8.3.6.3. Prospective binding mode reproduction.

Retrieval rates for the separate pharmacophores correspond to what is observed in the overlay of co-crystallized ligands with ~10% (7 compounds), 40-60% (30-45 compounds) and 50-70% (37-52) matching to pharmacophores 1, 2 and 3, respectively. The scoring of poses could however be improved for most algorithms, since more accurate poses are usually in the ensemble of solutions (**Figure 8.12C**) but are not scored as being best (**Figure 8.12A**). Although RMSD-based scoring methods (MOE, Pharmer, POT) have a comparable performance in pose prediction for Urokinase, they perform poorly in compound library enrichment (**Figure 8.13A,B**).

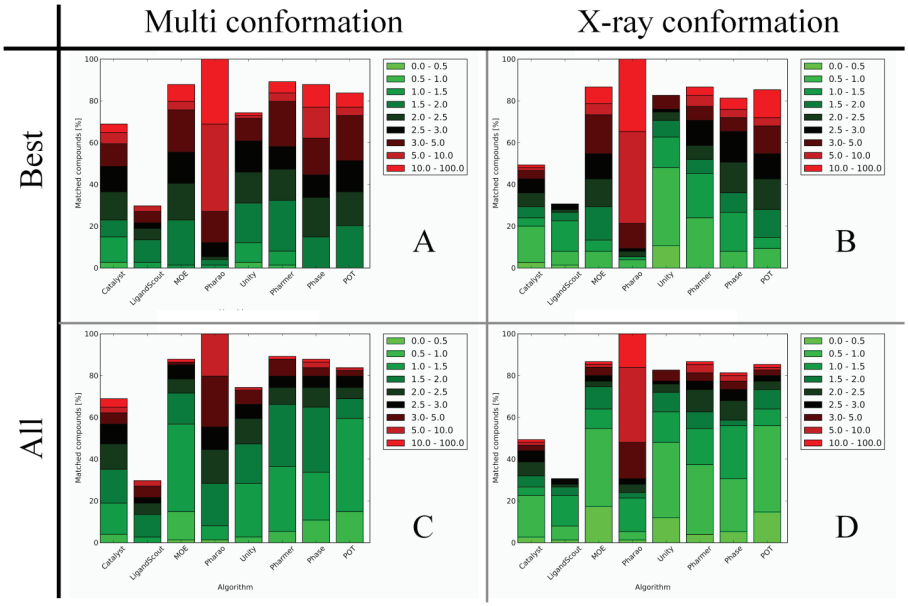
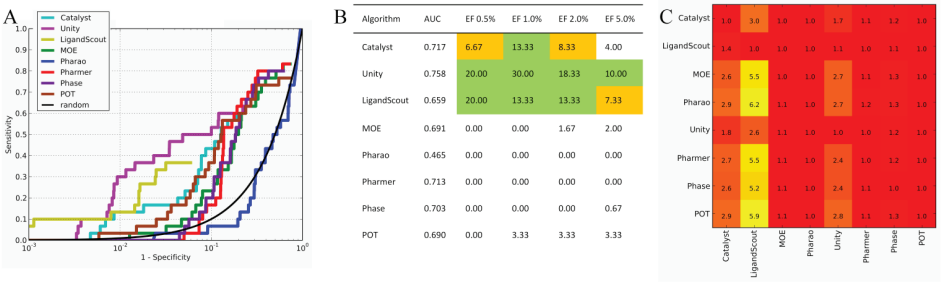


Figure 8.12: RMSD ranges of matched compounds from the co-crystallized ligand in two different scenarios: **A:** the best ranked pose from the ligand set in their multi-conformational format; **B:** the best ranked pose from the ligand set in X-ray conformation; **C:** lowest RMSD from all poses from the ligand set in multi-conformational format. **D:** lowest RMSD from all poses in the ligand set in their X-ray conformation.



Algorithm	AUC	EF 0.5%	EF 1.0%	EF 2.0%	EF 5.0%
Catalyst	0.717	6.67	13.33	8.33	4.00
Unity	0.758	20.00	30.00	18.33	10.00
LigandScout	0.659	20.00	13.33	13.33	7.33
MOE	0.691	0.00	0.00	1.67	2.00
Pharao	0.465	0.00	0.00	0.00	0.00
Pharmer	0.713	0.00	0.00	0.00	0.00
Phase	0.703	0.00	0.00	0.00	0.67
POT	0.690	0.00	3.33	3.33	3.33

Figure 8.13: Enrichment analysis of Urokinase MUV-dataset. **A:** ROC curves showing the enrichment of Urokinase actives/decoys (dataset created with MUV, see **Table 1**). **B:** AUC values and enrichment values at 0.5%, 1.0%, 2.0% and 5.0% false positive rate; green indicates an enrichment factor (EF) above 10.0 and orange indicates EF above 5.0. **C:** Heatmap that illustrates component importance when combining two pharmacophore algorithms.

8.3.6.4. Compound library enrichment.

Similarly to the other targets, overlay-based scoring algorithms outperform RMSD methods in compound library enrichment (**Fig 8.13A,B**). It is also notable that the combination of two knowledge based scoring algorithms (Catalyst, Unity and LigandScout) improves enrichment values (**Figure 8.13C**), while such trend is not observed for the combination of two RMSD based methods (**Figure 8.13C**). For instance, larger values in **Figure 8.13C**

are obtained if Catalyst, Unity and LigandScout are combined as consecutive screens with Pharmer, Phase, POT and MOE. The best enrichment is obtained by a combination of Catalyst with LigandScout: together both algorithms retrieve 10 out of 30 actives and 577 out of 15,000 decoys and have an enrichment factor of 8.7 (**paragraph 8.2.3.1.** eq.1).

8.4 General Discussion

The compound sets considered in this study have allowed us to explore the different characteristics of a range of pharmacophore screening algorithms in terms of compound retrieval and pose prediction. Different pharmacophores were derived from the ligand alignments observed in 80 CKD2, 120 CHK1, 110 PTP1B and 74 Urokinase crystal structures. For CDK2, 1 pharmacophore was defined from 45 actives, while respectively 36 and 21 actives defined 2 CHK1 pharmacophores, 21 and 77 actives defined 2 PTP1B pharmacophores and 7, 23 and 36 actives defined 3 Urokinase pharmacophores. Some of these pharmacophores match well-defined clusters of active molecules (CHK1 and PTP1B), while others match a more diverse range of chemical structures. Most algorithms retrieve a greater number of compounds than one would expect after analysis of the ligand poses in the available crystal structures, indicating that several active compounds are matched in conformations that do not correspond to the experimental one. The ability of the scoring functions to identify the correct bound pose is limited, as this can depend on i) the ligand input structures, ii) the pharmacophore's definition or iii) the scoring method applied by the pharmacophore screening tool. For instance, for CDK2, PTP1B and to a lesser extent CHK1, running the pharmacophore searches against only the X-ray conformations results in better pose reproductions than when searching against the conformational ensembles. The pharmacophore definition can also be responsible for poor pose reproduction as illustrated by the PTP1B pharmacophore 2, which describes only a small part of the interaction patterns of the co-crystallized ligands and generates only poses with relatively high RMSD values. Scoring methods are also sub-optimal, as they frequently fail to identify the best pose from the full ensemble of matched poses (see CDK2 and Urokinase datasets). Compound retrieval based on poses dissimilar to the biophysical binding mode suggests that hit identification by serendipity does still frequently occur in pharmacophore search strategies.

Compound library enrichment seems not so much related to the pharmacophore search algorithm used but more by the compound sets and corresponding pharmacophores used in this study. PTP1B, particularly, showed very good early enrichments that might be related to the N-substituted oxamic acid moiety present in most active compounds.

Combining the strength of several algorithms seems possible by screening compound libraries with different algorithms in a consecutive order. By comparison of the enrichment factor of all compounds matching algorithm pairs and enrichment factor of all compounds matching a single algorithm, we showed that improvements over a factor of 1.5 are possible.

8.5 Conclusions

We carried out a comparative study of eight pharmacophore screening tools for their ability to retrieve and describe the behavior of active compounds for four biological targets of interest. Several analyses allowed us to better elucidate advantages and drawbacks of algorithms when they are applied with their default settings for high-throughput pharmacophore screening purposes.

Our analysis shows that the correct reproduction of experimental binding poses is generally better with algorithms using RMSD-based methods (MOE, Pharmer, Phase and POT) than with overlay-based methods (Catalyst, LigandScout, Pharao and Unity). However, since many compounds match pharmacophore hypotheses without reproducing the experimental pose, it is also important to assess the ratio of correctly predicted compounds on the overall number of matched compounds in a given data set. In this respect, the performance of overlay-based algorithms is slightly better than the RMSD-based methods. Thus, while RMSD-based algorithms generally return 'more shots on goal' due to the high number of poses, overlay-based methods seem to provide the best chance of retrieving the relevant biophysical binding mode. Taken as a whole, these findings suggest that one may prefer certain algorithms depending on the research application. For example, when optimizing lead compounds, it may be necessary to collect a large number of binding modes to explore the conformational space thoroughly. In such a case one may prefer an RMSD-based method. Vice versa, if only a single binding mode is desired, one may prefer overlay-based methods. In fact, the stricter criteria of overlay-based methods return better results in compound library enrichment studies, as they are better at discriminating between active and inactive compounds. As this is most likely due to the stricter fitting criteria (which retrieve subsets of RMSD-based methods) and better scoring, it seems feasible to pre-screen large compound databases with RMSD-based pharmacophore screening methods, which are typically faster, to obtain the same results in a less time-consuming manner.

Overall we can conclude that: i) the more advanced overlay-based scoring algorithms have better enrichments ii) obtaining good enrichment is dependent on the biological target and not strictly on the choice of a given algorithm iii) the use of different pharmacophore search algorithms may lead, in their default settings, to non-negligible different results.

We acknowledge that our findings could be extended and corroborated with further analyses with other biological targets, but the findings of this paper may be of practical use in the planning of more efficient high-throughput pharmacophore screens.

REFERENCES

1. Schneider, G., K. Baringhaus, and H. Kubinyi, *Molecular Design: Concepts and Applications*. 2008, Weinheim: Wiley-VCH.
2. Varnek, A. and A. Tropsha, *Chemoinformatics Approaches to Virtual Screening*. 2008, London: RCS.
3. Gasteiger, J. and T. Engel, *Chemoinformatics*. 2003, Weinheim: Wiley-VCH.
4. Caporuscio, F., et al., *Structure-Based Design of Potent Aromatase Inhibitors by High-Throughput Docking.* J Med Chem, 2011. **54**(12): p. 4006–17.
5. Kolb, P., et al., *Docking and chemoinformatic screens for new ligands and targets.* Current Opinion in Biotechnology, 2009. **20**(4): p. 429-36.
6. Villoutreix, B.O., et al., *Free resources to assist structure-based virtual ligand screening experiments.* Curr Protein Pept Sci, 2007. **8**(4): p. 381-411.
7. Ballester, P.J., et al., *Prospective virtual screening with Ultrafast Shape Recognition: the identification of novel inhibitors of arylamine N-acetyltransferases.* Journal of the Royal Society Interface, 2010. **7**(43): p. 335-42.
8. Schneider, G., *Virtual screening: an endless staircase?* Nat Rev Drug Discov, 2010. **9**(4): p. 273-6.
9. Leach, A.R., et al., *Three-dimensional pharmacophore methods in drug discovery.* J Med Chem, 2010. **53**(2): p. 539-58.
10. Langer, T., *Pharmacophores in Drug Research.* Molecular Informatics, 2010. **29**(6-7): p. 470-5.
11. Caporuscio, F. and A. Tafi, *Pharmacophore modelling: a forty year old approach and its modern synergies.* Current Medicinal Chemistry, 2011. **18**(17): p. 2543-53.
12. Sun, H., *Pharmacophore-based virtual screening.* Current Medicinal Chemistry, 2008. **15**(10): p. 1018-24.
13. Del Rio, A., A. Barbosa, and F. Caporuscio, *Use of large multiconformational databases with structure-based pharmacophore models for fast screening of commercial compound collections.* Journal of Cheminformatics, 2011. **3**(Suppl 1): p. P27.
14. Del Rio, A., et al., *CoCoCo: a free suite of multiconformational chemical databases for high-throughput virtual screening purposes.* Molecular Biosystems, 2010. **6**(11): p. 2122-8.
15. Irwin, J.J. and B.K. Shoichet, *ZINC--a free database of commercially available compounds for virtual screening.* J Chem Inf Model, 2005. **45**(1): p. 177-82.
16. Masciocchi, J., et al., *MMsINC: a large-scale chemoinformatics database.* Nucleic Acids Res, 2009. **37**(Database issue): p. D284-90.
17. Barbosa, A.J. and A. Del Rio, *Freely accessible databases of commercial compounds for high-throughput virtual screenings* Current Topics in Medicinal Chemistry, 2011. **Accepted**.
18. Kolb, P. and J.J. Irwin, *Docking screens: right for the right reasons?* Curr Top Med Chem, 2009. **9**(9): p. 755-70.
19. Chen, Z., et al., *Pharmacophore-based virtual screening versus docking-based virtual screening: a benchmark comparison against eight targets.* Acta Pharmacologica Sinica, 2009. **30**(12): p. 1694-708.
20. Kirchmair, J., et al., *Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection--what can we learn from earlier mistakes?* J Comput Aided Mol Des, 2008. **22**(3-4): p. 213-28.
21. ten Brink, T. and T.E. Exner, *Influence of protonation, tautomeric, and stereoisomeric states on protein-ligand docking results.* J Chem Inf Model, 2009. **49**(6): p. 1535-46.
22. Patel, Y., et al., *A comparison of the pharmacophore identification programs: Catalyst, DISCO and GASP.* J Comput Aided Mol Des, 2002. **16**(8-9): p. 653-81.
23. Peach, M.L. and M.C. Nicklaus, *Combining docking with pharmacophore filtering for improved virtual screening.* Journal of Cheminformatics, 2009. **1**(1): p. 6.
24. Brown, S.P. and S.W. Muchmore, *Large-scale application of high-throughput molecular mechanics with Poisson-Boltzmann surface area for routine physics-based scoring of protein-ligand complexes.* J Med Chem, 2009. **52**(10): p. 3159-65.
25. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
26. Morgan, D.O., *Cyclin-dependent kinases: engines, clocks, and microprocessors.* Annu Rev Cell Dev Biol, 1997. **13**: p. 261-91.
27. Sanchez, Y., et al., *Conservation of the Chk1 checkpoint pathway in mammals: linkage of DNA damage to Cdk regulation through Cdc25.* Science, 1997. **277**(5331): p. 1497-501.

28. Elchebly, M., et al., *Increased insulin sensitivity and obesity resistance in mice lacking the protein tyrosine phosphatase-1B gene.* Science, 1999. **283**(5407): p. 1544-8.
29. Andreasen, P.A., et al., *The urokinase-type plasminogen activator system in cancer metastasis: a review.* Int J Cancer, 1997. **72**(1): p. 1-22.
30. Apweiler, R., et al., *UniProt: the Universal Protein knowledgebase.* Nucleic Acids Res, 2004. **32**(Database issue): p. D115-9.
31. Ulbert, I., et al., *In vivo laminar electrophysiology co-registered with histology in the hippocampus of patients with temporal lobe epilepsy.* Experimental Neurology, 2004. **187**(2): p. 310-8.
32. *Epik*, 2011, Schrödinger, LLC.
33. Renner, S. and G. Schneider, *Fuzzy pharmacophore models from molecular alignments for correlation-vector-based virtual screening.* J Med Chem, 2004. **47**(19): p. 4653-64.
34. Wild, D.J. and C.J. Blankley, *Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using Ward's clustering.* J Chem Inf Comput Sci, 2000. **40**(1): p. 155-62.
35. *ChEMBL*. 2011; Available from: https://www.ebi.ac.uk/chembldb/.
36. Warr, W., *ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI).* J Comput-Aided Mol Des, 2009. **23**(4): p. 195-8.
37. Rohrer, S.G. and K. Baumann, *Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data.* J Chem Inf Model, 2009. **49**(2): p. 169-84.
38. Boot, W.R., et al., *Automatic and intentional memory processes in visual search.* Psychon Bull Rev, 2004. **11**(5): p. 854-61.
39. Watts, K.S., et al., *ConfGen: a conformational search method for efficient generation of bioactive conformers.* J Chem Inf Model, 2010. **50**(4): p. 534-46.
40. Ku, C.H., et al., *Large-scale gene expression analysis of osteoblasts cultured on three different Ti-6Al-4V surface treatments.* Biomaterials, 2002. **23**(21): p. 4193-202.
41. Boulkroun, S., et al., *Vasopressin-inducible ubiquitin-specific protease 10 increases ENaC cell surface expression by deubiquitylating and stabilizing sorting nexin 3.* Am J Physiol Renal Physiol, 2008. **295**(4): p. F889-900.
42. Koes, D.R. and C.J. Camacho, *Pharmer: efficient and exact pharmacophore search.* J Chem Inf Model. **51**(6): p. 1307-14.
43. Sanders, M.P., et al., *Snooker: A Structure-Based Pharmacophore Generation Tool Applied to Class A GPCRs.* J Chem Inf Model, 2011. **51**(9): p. 2277-92.
44. Taminau, J., G. Thijs, and H. De Winter, *Pharao: pharmacophore alignment and optimization.* J Mol Graph Model, 2008. **27**(2): p. 161-9.
45. Claverol, E.T., A.D. Brown, and J.E. Chad, *A large-scale simulation of the piriform cortex by a cell automaton-based network model.* IEEE Transactions on Biomedical Engineering, 2002. **49**(9): p. 921-35.
46. Wolber, G. and T. Langer, *LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters.* J Chem Inf Model, 2004. **45**(1): p. 160-9.
47. Wolber, G. and T. Langer, *LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters.* J Chem Inf Model, 2005. **45**(1): p. 160-9.
48. *Phase*, 2011, Schrödinger, LLC.
49. Zou, J., et al., *Towards more accurate pharmacophore modeling: Multicomplex-based comprehensive pharmacophore map and most-frequent-feature pharmacophore model of CDK2.* J Mol Graph Model, 2008. **27**(4): p. 430-8.
50. Agrafiotis, D.K., *Stochastic proximity embedding.* J Comput Chem, 2003. **24**(10): p. 1215-21.
51. Agrafiotis, D.K. and H. Xu, *A self-organizing principle for learning nonlinear manifolds.* Proc Natl Acad Sci U S A, 2002. **99**(25): p. 15869-72.
52. Ignatovich, O., et al., *The creation of diversity in the human immunoglobulin V(lambda) repertoire.* J Mol Biol, 1997. **268**(1): p. 69-77.
53. Chen, X.M., et al., *Structure-based and shape-complemented pharmacophore modeling for the discovery of novel checkpoint kinase 1 inhibitors.* J Mol Model, 2010. **16**(7): p. 1195-204.
54. Zhang, S. and Z.Y. Zhang, *PTP1B as a drug target: recent developments in PTP1B inhibitor discovery.* Drug Discov Today, 2007. **12**(9-10): p. 373-81.
55. Shui, L., et al., *Urokinase Inhibitor Design Based on Pharmacophore Model Derived from Diverse Classes of Inhibitors.* IBC, 2009.

Summary

Summary

The output of the global drug discovery efforts has been very disappointing the last years. Despite ever increasing investments in research and development there has not been an increase in the number of approved drugs. To make matters worse, the patents of many blockbuster drugs will expire in the coming few years . To prevent a stagnation of the drug industry, new drug candidates with improved properties are required. This thesis describes a molecular class-specific information system which contains large amounts of heterogeneous data on G protein-coupled receptors (GPCRs) as well as a method to predict the ligand interacting residues from these data and to translate them into pharmacophore features which describe ligand features complementary to these residues. These pharmacophores can either be used to identify new chemical entities which show activity on the target or help to understand the relationship between ligand structures and their activity on the protein target. **Chapter 2** reviews the methodologies and software tools which are available for structure based pharmacophore modeling. Methods to derive pharmacophore features typically use the geometric interaction properties of residues or observed feature locations of ligands interacting with the protein. The selection of the features which are essential for biological activity can be based on either interaction energies or experimental data. **Chapter 3** describes a GPCR specific information system (GPCRdb) containing experimental data on sequences, ligand binding constants, mutations, and oligomers, as well as many different types of computationally derived data such as multiple sequence alignments and homology models. As such, the GPCRdb is a good starting point for drug discovery programs targeted at GPCRs. In **chapter 4** we report about a method to predict ligand-interacting residues located in the transmembrane domains of GPCRs. The intracellular signaling cascade for GPCRs is evolutionary very successful as witnessed by the limited number of intracellular signaling pathways. In contrast, the endogenous ligands do vary between different subfamilies and show only similarity within small subfamilies. Based on this we hypothesized that ligand binding residues are conserved in a multi species multiple sequence alignment (MSA) of the members of a small subfamily and not in a MSA of the entire class A GPCR family. We proved that we are indeed capable of selecting ligand interacting residues after assessing a measure of sequence conservation for both MSA's. **Chapter 5** describes a method which uses the identified ligand interacting residues to generate pharmacophores. These pharmacophores describe desired ligand features in the receptor ligand binding pockets and are used for binding mode hypotheses generation and compound library enrichment. We retrospectively show that our method is able to reproduce literature supported binding modes for the β2 adrenergic receptor. Since most endogenous GPCR ligands are agonists it is likely that the ligand binding residue predictions and eventually resulting pharmacophores are biased to agonism. This hypothesis is in agreement with the observation that the retrieval and prediction of agonists is better than for antagonists with our β2 adrenergic receptor pharmacophore. Furthermore we showed

that our structure based pharmacophores were able to generate enriched target specific compound libraries for several different GPCRs. In chapter 6 and 7 we present the results of two prospective experiments. **Chapter 6** reviews the outcome of the international community-wide assessment of GPCRdock structure modeling and ligand docking assessment. Binding mode hypotheses of eticlopride in the dopamine D3 receptor based on our pharmacophore modeling tool and optimized with the flexible docking tool Fleksy scored 2nd , 6th, 7th, 9th and 10th out of over 100 submitted predictions. **Chapter 7** describes the application of pharmacophores in combination with a frequent substructure mining technique of known actives in compound library selection. New active compounds for an adenosine A2a, β2 adrenergic and sphingosine 1-phosphate receptor were successfully identified from the selected libraries. In chapter 8 we investigated the performance of several pharmacophore search algorithms on ligand binding mode reproduction and compound library enrichment. From the fact that a fraction of the active molecules was predicted to be active based on an incorrect binding mode , we can conclude that the discovery of new active molecules by means of pharmacophore techniques still happens partly due to serendipity .

Samenvatting

Samenvatting

De laatste decennia is er in toenemende mate geïnvesteerd in onderzoek naar en ontwikkeling van medicijnen. Maar desondanks zijn er veel minder nieuwe chemische entiteiten (stoffen) als medicijn op de markt gebracht als aanvankelijk geanticipeerd. Van vele succesvolle medicijnen lopen de octrooien af en dus ook de bescherming van de markt. Dit alles leidt er toe dat de farmaceutische industrie in zwaar weer is komen te zitten. Er zijn dringend nieuwe medicijnen nodig, in de eerste plaats om de patiënten te helpen en in de tweede plaats om de medicijnindustrie in staat te stellen te investeren in research naar meer en betere medicijnen.

Dit proefschrift beschrijft een informatiesysteem dat een grote hoeveelheid gegevens bevat, afkomstig van verschillende bronnen. De gegevens zijn specifiek voor een familie van eiwitten, genaamd G-eiwit gekoppelde receptoren (GPCR). Verder wordt een methode beschreven die het mogelijk maakt te voorspellen welke aminozuren van cruciaal belang zijn voor ligand-eiwitinteracties en een methode om de informatie van deze aminozuren te vertalen in pharmacophoren die de gewenste ligandeigenschappen in de context van het eiwit beschrijven (een pharmacophore is een abstracte beschrijving van moleculaire eigenschappen aan welke een ligand herkend kan worden door een eiwit.). Deze op eiwitstructuur gebaseerde pharmacophoren kunnen gebruikt worden om nieuwe stoffen te vinden die effect zullen hebben op het betreffende eiwit. Ook kunnen deze pharmacophoren gebruikt worden om activiteiten van reeds bekende stoffen te verklaren.

In **Hoofdstuk 2** worden een aantal methodes en softwareapplicaties beschreven die beschikbaar zijn voor het genereren van op een eiwitstructuur gebaseerde pharmacophoren. Deze methodes gebruiken de typisch geometrische kenmerken van eiwit-ligandinteracties of de daadwerkelijk waargenomen ligandkarakteristieken in eiwit-ligandcomplexen, om de plaatsen te bepalen waar een voorkeur is voor een bepaalde feature of eigenschap van een ligand. De uiteindelijke selectie van eigenschappen die cruciaal zijn voor de biologische activiteit leidt tot een pharmacophore en is vaak gebaseerd op experimentele data of indien niet aanwezig op de berekening van interactie-energieën.

Hoofdstuk 3 beschrijft een GPCR specifiek informatiesysteem (GPCRdb) met experimentele data over aminozuursequenties, bindingsconstanten van liganden, aminozuurmutaties, en oligomeren, alsmede verschillende modellen en data afgeleidt van multiple sequence alignments en homologiemodellen. Hiermee is de GPCRdb een goed beginpunt voor medicijnonderzoekprogramma's die gericht zijn op GPCRs.

Hoofdstuk 4 behandelt een methode die gebruikt kan worden om aminozuren te voorspellen in de transmembraandomeinen van GPCRs die betrokken zijn bij ligandinteracties. De methode maakt gebruik van de observering dat de intracellulaire signaaltransductie slechts plaatsvindt via een beperkt aantal G-eiwitten, terwijl de natuurlijke liganden varïeren tussen verschillende subfamilies maar ook grote overeenkomsten vertonen in

de verschillende subfamilies. Gebaseerd op deze observering veronderstellen wij dat ligandbindende aminozuren geconserveerd zijn in de multiple sequence alignment (MSA) van een kleine subfamilie van GPCRs van een groot aantal verschillende organismen en dat dit niet het geval is in de MSA van de hele class A GPCR-familie. Doormiddel van een berekening van de conservering van aminozuurposities in een MSA van verschillende subfamilies en van de MSA van alle class A GPCRs bewijzen we dat we inderdaad met deze method in staat zijn om ligandbindende aminozuren te identificeren.

In **hoofdstuk 5** wordt vervolgens een methode gepresenteerd die de geïdentificeerde ligandbindende aminozuren gebruikt om pharmacophoren te genereren. Deze pharmacophoren beschrijven gewenste ligandeigenschappen in de ligandbindingsholte van een receptor en worden gebruikt om hypotheses af te leiden voor bekende liganden en voor het ontwerpen van bibliotheken met chemische stoffen. In dit hoofdstuk laten we met behulp van een retrospectieve studie zien dat de gepresenteerde methode in staat is pharmacophoren te genereren waarmee de bindingsmodes van β2-adrenerge stoffen in de β2-adrenerge-receptor kunnen worden gereproduceerd. Aangezien de meeste natuurlijke liganden van GPCRs agonisten zijn is het waarschijnlijk dat de ligandbindende aminozuurvoorspellingen en de hieruit voortkomende pharmacophoren gericht zijn op agonisme. Dit verklaart wellicht ook de observering dat agonisten beter herkend worden door de pharmacophore voor de β2-adrenerge-receptor. Verder laten we in dit hoofdstuk ook zien dat gegenereerde pharmacophoren in staat zijn om stoffenbibiliotheken te maken die een groter percentage actieve stoffen bevat tegen de receptor waarvan de pharmacophore is afgeleid dan stoffenbibiliotheken die met een pharmacophore voor een andere receptor zijn gemaakt.

In **hoofdstuk 6** en **7** presenteren we twee voorspellende experimenten. **Hoofdstuk 6** beschrijft de uitkomsten van GPCR-structuurmodellering en ligandplaatsing, in het kader van een evaluatie die plaatsvond in de internationale gemeenschap als een soort wedstrijd voordat de kristalstructuren bekend werden gemaakt. Voorspellingen van de ligandbindingsmodus van eticlopride in de dopamine-D3-receptor gebaseerd op pharmacophoren afgeleidt met de methode beschreven in hoofdstuk 5 en geoptimaliseerd met het flexibele dockingprogramma Fleksy eindigden op de 2de, 6de, 7de, 9de en 10de plaats uit in totaal meer dan 100 inzendingen.

In **hoofdstuk 7** wordt beschreven hoe pharmacophoren in combinatie met een mining (zoekstrategie) techniek gebaseerd op veel voorkomende substructuren in bekende liganden wordt gebruikt om een kleine stoffenbibliotheek samen te stellen uit een grote collectie van stoffen. Nieuwe stoffen met activiteit op de adenosine-A2a, β2-adrenerge en sphingosine 1-fosfaat receptor bleken aanwezig te zijn in de geselecteerde stoffen-bibliotheek.

Hoofdstuk 8 onderzoekt de kwaliteit van verschillende algorithmen die stoffenbibliotheken doorzoeken met behulp van pharmacophoren. Kwaliteitscriteria die in dit hoofdstuk zijn beschreven omvatten de reproductie van ligandbindingsmodi en de mate van verrijking

gevonden door stoffenbibliotheken te filteren op stoffen die passen in de pharmacophore. Gebaseerd op het feit dat een deel van de bekende actieve stoffen voorspeld werd met een bindingsmodus die niet overeenkomt met de bio-actieve bindingsmodus kan geconcludeerd worden dat de ontdekking van nieuwe actieve stoffen met behulp van pharmacophore technieken nog steeds deels plaatsvindt doormiddel van toeval.

Dankwoord

Dit proefschrift duidt voor mij het einde aan van een leerzame en bijzondere periode. Al dit werk heeft echter niet tot stand kunnen komen zonder de hulp van een grote groep mensen om mij heen. Zowel begeleiders, collega's als vrienden hebben mij gedurende deze vier jaar geholpen met adviezen met betrekking tot mijn werk en leven en hebben gezorgd voor de ontspanning die nodig zijn om een proefschrift te schrijven.
Natuurlijk was dit alles niet mogelijk zonder de ondersteuning van mijn promotor (Jacob de Vlieg) en begeleider (Jan Klomp). Als nog bachelor student kwam ik Jacob tegen bij de cursus "Bioinformatics for Drug Discovery'' in Nijmegen. Dit deed mij besluiten om uiteindelijk ook voor mijn afstudeerproject te werken bij de afdeling Moleculer Design & Informatics (MDI) bij Organon NV. Hier kwam ik op de kamer te zitten bij Jan Klomp. Mijn afstudeerproject was mede hierdoor erg leuk met vele zowel inhoudelijke als persoonlijke gesprekken. Toen ik op zoek was naar een promotieplaats en jullie mij deze aanboden heb ik daarom ook niet lang getwijfeld. Ik ben jullie dan ook erg dankbaar met jullie vertrouwen in mij en dat jullie mij de kans gegeven hebben om mijn promotie onder jullie leiding te voltooien. Ik heb gedurende deze tijd erg veel van jullie geleerd op verschillende vlakken.

Jacob, ik heb erg veel bewondering voor de inzet die jij hebt getoond voor de integratie van computer modellen in het onderzoek naar nieuwe geneesmiddelen. Dat het in de ontwikkeling van software niet slechts draait om het maken van 'leuke' tools, maar dat het vooral om de toepasbaarheid wordt vaak vergeten. Jouw constante focus op deze toepasbaarheid heeft mij erg geholpen en heeft mij doen beseffen dat onderzoek teamwork is waarin het cruciaal is dat mensen met verschillende achtergronden met elkaar kunnen discussiëren en dat ontwikkelde software hierop aangepast dient te worden.

Jan, jouw enthousiasme en wetenschappelijke nieuwsgierigheid heeft mij heel gemotiveerd en gestimuleerd wat uiteindelijk geresulteerd heeft in dit proefschrift. Ik zal de vele gesprekken op jou kamer voor het whiteboard niet snel vergeten. Ik vond het altijd erg fijn om samen te brainstormen over mogelijke oplossingen van problemen en om gewoon even bij elkaar binnen te lopen voor een praatje. Jij hebt een grote invloed gehad op mijn ontwikkeling van master student tot doctor ingenieur.

Natuurlijk wil ik ook graag alle mensen bij MDI bedanken voor de fijne en leerzame tijd. Werken in een prettige en goed gestructureerde omgeving waarin databases op orde zijn, software werkt zoals dit hoort en mensen gebruik maken van nieuwe ontwikkelingen en feedback geven zorgen uiteindelijk voor een beter eind resultaat. Zodra je een aantal namen noemt is het bijna zeker dat je er ook een aantal vergeet, maar zonder de hulp van Stefan, Tinka, Ruud, Ross, Scott, Markus, Jos, Ria, Thea en Karin, en natuurlijk Hans was

het zeker niet gelukt. Stefan, heel erg bedankt voor al je hulp in mijn project. Het op orde houden van onze database en de ontwikkeling van webinterface hebben geresulteert in een prachtig eindproduct. Tinka en Ruud, heel erg bedankt voor het installeren van alle software en het oplossen van alle computer gerelateerde problemen. Markus en Jos, heel erg bedankt dat ik al jullie programma'tjes en scriptjes mocht gebruiken en de vele project gerelateerde adviezen. Ross en Scott, bedankt voor alle adviezen en natuurlijk het nakijken van al mijn geschreven engels. De secrateresses, Ria, Thea en Karin, ook jullie wil ik graag bedanken voor jullie hulp. Jullie waren er altijd als ik iets nodig had of niet wist hoe ik bepaalde dingen moest regelen. En als laatste natuurlijk Hans. Jij was gedurende vier jaar mijn kamergenoot. Ik heb altijd erg met je gelachen. Ondanks dat we samen wat problemen hadden met het op orde houden van onze kamer en het invullen van ons labjournaal vond ik het erg fijn om met jou een kamer delen en heb ik ontzettend veel lol gehad tijdens deze periode.

Ook moet ik zeker mijn collega's binnen de computational drug design (CDD) groep niet vergeten in dit dankwoord. Samen hebben we denk ik een leuke tijd gehad in Oss ook al was het met alle overnamens ook voor ons niet altijd makkelijk. Allereerst Sander, onder jouw leiding heb ik mijn afstudeerproject voltooid en jij stond ook tijdens mijn promotie altijd klaar om vragen voor me te beantwoorden, brainstormen en om mij te leren hoe ik eigenlijk een artikel moest schrijven. Ik heb erg veel van je geleerd en ben ook heel erg blij dat je mijn co-promotor wil zijn voor dit proefschrift. Tina en Dave, jullie waren de andere 'modellers' in het groepje. Naast jullie input voor mijn project en het lezen van mijn manuscripten ben ik jullie ook erg dankbaar voor de gezellige tijd en vele koffie'tjes om 10 en 3 uur. Dave, het samen lopen van de halve marathon had ik waarschijnlijk ook nooit zonder jou gedaan. Jammer alleen dat we elkaar kwijt raakten in de drukte en ik ietsje trager was. Hierdoor moeten we dit misschien toch nog maar eens overdoen. Verder wil ik ook de andere mensen binnen CDD bedanken. Wilco, Raoul en Eugene, heel erg bedankt voor de discussies en de gezelligheid.

Verder zijn er dan natuurlijk nog de studenten die ik heb mogen begeleiden. Sven, jou werk heeft mede geleidt tot hoofdstuk 4 van dit boekje. Ik heb heel fijn met je samengewerkt en veel geleerd in ons kleine projectteam met ook Jan en Stefan. Cizar en Gwen, ook bedankt voor jullie werk en dat ik jullie heb mogen begeleiden tijdens jullie stages.
Door de sluiting van de onderzoeksafdelingen van MSD in Oss, heb ik de laatste maanden van mijn promotie gewerkt op de Radboud Universiteit in Nijmegen en de Vrije Universiteit van Amsterdam. Ik ben erg welkom geweest op beide werkplekken en wil graag de mensen bedanken die dit mogelijk gemaakt hebben. Barbara heel erg bedankt met alle hulp omtrend alle administratieve zaken omtrend mijn promotie.

Ondanks dat ik geen promovendus was aan de VU in Amsterdam was ik hier wel erg welkom en ik wil daarom graag de afdeling medicinal chemistry bedanken voor hun gastvrijheid. In het bijzonder de mensen die mij geholpen hebben met het schrijven van verschillende artikelen en afronden van mijn proefschrift. Luc, ik vond het erg fijn om samen met jou een team te vormen wat met name gezorgd heeft voor de hoofdstukken 2, 6 en 7. Chris ook jij heel erg bedankt voor al jou hulp bij het schrijven. Helaas is dit niet mijn sterkste en ik ben dan ook zeer dankbaar met jou hulp. Dit is zonder twijfel heel erg belangrijk geweest voor de tot stand koming van dit proefschrift. Als laatste wil ik ook graag aan mijn overige roommates op de VU bedanken. Albert en Dana, ik vond het erg fijn een bureau bij jullie op de kamer te hebben en heb zeker meer geleerd over het leven (en dan vooral over hoe dit buiten Brabant is).

Zoals iedereen wel weet is succesvol zijn in je werk niet mogelijk zonder de steun van vrienden en familie. De vele biertjes, schuine moppen en gezellige avonden samen met al mijn vrienden hebben mij erg geholpen om het werk even los te laten en gewoon te genieten om hierna op maandag erweer tegenaan te gaan. Heel erg bedankt hiervoor en ik hoop dat we nog vele biertjes samen kunnen drinken en gezellige avonden mogen hebben.

Als laatste wil ik dan heel graag mijn familie bedanken. Allereerst mijn broer Erik, samen hebben we tijdens jouw promotie heel wat tijd samen doorgebracht in de auto naar Eindhoven. We hebben vaak gepraat over promoveren. Met jou kon ik altijd goed van gedachte wisselen. Je bent voor mij naast mijn broer dan ook een heel erg goede vriend. Mijn schoonzus, Rinske, jij bent inmiddels ook begonnen aan je promotie en met jou heb ik de laatste paar maanden regelmatig gecarpoold naar Nijmegen. Bedankt voor deze gezellige ritjes samen. En natuurlijk mijn ouders. Jullie hebben mij altijd gesteund in mijn studie en promotie. Met ‘ons pap’ ben ik de eerste jaren samen naar Oss gereden en heb ik veel gesprekken gevoerd over het bedrijfsleven, de pharmaceutische industrie en management. Ik heb je gedurende deze jaren heel goed leren kennen en ben je erg dankbaar voor deze gesprekken. ‘Ons ma’ die er altijd voor mij was, elke dag vroeg hoe het ging en ons stimuleerde om met onze studie een goede basis te leggen voor het verdere leven. Als allerlaatste dan ‘mijn meisje’. Joey, ik ben erg blij dat ik je tegen gekomen ben en voor jouw steun tijdens de laatste jaren van mijn promotie. Het was voor mij niet altijd makkelijk met de vele veranderingen op mijn werk, maar jij hielp mij altijd te relativeren. Het allerbelangrijkste is immers dat je werk doet wat je leuk vind en een leuk leven hebt. Ik hoop dat wij nog lang samen van het leven mogen genieten.

Nogmaals dank aan iedereen die mij heeft gesteund!

Marijn

Curriculum vitae

Marijn Sanders werd op 6 Juli 1984 geboren te Uden. Na het behalen van het V.W.O.-diploma aan het Udens College te Uden in 2002, begon hij aan de studie Biomedische Technologie aan de Technische Universiteit Eindhoven (TU/e). Gedurende zijn bachelor fase heeft hij tevens meegewerkt bij het Epilepsie centrum Kempenhaeghe aan de ontwikkeling van een database voor de detectie van epilepsie-aanvallen. Tijdens de masterfase van zijn opleiding heeft hij in de vorm van een interne stage onder leiding van dr. ir. L. Roumen en prof. dr. P. Hilbers onderzoek verricht aan moleculaire simulaties van CYP11B-ligandcomplexen. Na deze stage is hij begonnen aan zijn afstudeerproject, dat een samenwerking was tussen de afdeling Molecular Design & Informatics (MDI) van Organon NV en de afdeling Biomodeling en Bioinformatics van de TU/e. Onder leiding van dr. S. Nabuurs, dr. M. Wagener, dr. ir. K. Pieterse, prof. dr. J. de Vlieg en prof. dr. P. Hilbers werkte hij hiervoor aan de evaluatie van de betrouwbaarheid van fragmentdocking en een analyse van relevante factoren hiervoor. Hij heeft in die periode ook nog vier maanden aan de Universiteit van Uppsala onder leiding van dr. D. van der Spoel onderzoek gedaan naar waterstofbrugpatronen in eiwitten. Hij haalde in 2007 zijn masterdiploma biomedisch ingenieur aan de TU/e. Naast zijn studie heeft hij tevens met succes het certificaat technisch management behaald.

Hierna is hij begonnen aan een onderzoeksproject in een samenwerkingsverband tussen de afdeling MDI van Organon NV in Oss en de Computational Drug Discovery (CDD)-groep van het Radboud Universitair Medisch Centrum te Nijmegen onder auspiciën van het Top Instituut Pharma. Hij ontwikkelde een software methode, waarmee op eiwitstructuur gebaseerde pharmacophoren kunnen worden gegenereerd. Deze techniek is vervolgens in samenwerkingsverbanden met onder andere de Universiteiten van Leiden, Amsterdam, Maastricht en Bologna toegepast in verschillende projecten. Gedurende het grootste gedeelte van het onderzoek is hij werkzaam geweest op de afdeling MDI van Organon N.V.in Oss onder leiding van dhr. J. Klomp en prof. dr. J. de Vlieg. De resultaten van het onderzoek zijn beschreven in dit proefschrift.

Bibliography

Sanders MP, McGuire, R, Roumen L, de Esch IJ, de Vlieg J, Klomp JP, de Graaf C, **From the protein's perspective: The benefits and challenges of protein structure-based pharmacophore modelling**, 2011, MedChemComm, accepted

Vroling B, Sanders M, Baakman C, Borrmann A, Verhoeven S, Klomp J, Oliveira L, de Vlieg J, Vriend G, **GPCRDB: information system for G protein-coupled receptors**, Nucleic Acids Res, 2011, 39(database issue), D309-19

Sanders MP, Fleuren WW, Verhoeven S, van den Beld S, Alkema W, de Vlieg J, Klomp JP, **ss-TEA: Entropy based identification of receptor specific ligand binding residues from a multiple sequence alignment of class A GPCRs**, BMC Bioinformatics, 2011, 12, 322

Sanders MP, Verhoeven S, de Graaf C, Roumen L, Vroling B, Nabuurs SB, de Vlieg J, Klomp JP, **Snooker: A structure based pharmacophore generation tool applied to class A GPCRs**, J Chem Inf Model, 2011, 51(9), p. 2277-92

Roumen L, Sanders MP, Vroling B, de Esch IJ, de Vlieg JP, Leurs R, Klomp JP, Nabuurs SB, de Graaf C, ***In silico veritas:*** **The pitfalls and challenges of predicting GPCR-ligand interactions**, 2011, Pharmaceuticals, 4(9), 1196-1215

Sanders MP, Roumen L, van der Horst E, Lane JB, Vischer HF, van Offenbeek J, de Vries H, Verhoeven S, Chow KY, Verkaar F, Beukers MW, McGuire R, Leurs R, IJzerman AP, de Vlieg J, de Esch IJ, Zaman GJ, Klomp JP, de Graaf C, Bender A, **A prospective cross-sceening study on G protein-coupled receptors: Lessons learned in virtual compound library design**, 2011, in preparation

Sanders MP, Barbosa AJ, Zarzycka B, Nicolaes GA, Klomp JP, de Vlieg J, Del Rio A, **A comparative analysis of pharmacophore screening tools**, 2011, J Chem Inf Model, submitted

Roumen L, Sanders MP, Pieterse K, Hilbers PA, Plate R, Custers E, de Gooyer M, Smits JF, Beugels I, Emmen J, Ottenheijm HC, Leysen D, Hermans JJ, **Construction of 3D models of the CYP11B family as a tool to predict ligand binding characteristics**, J Comput Aided Mol Des, 2007, 21(8), p. 455-71